Journal of Data Science 12(2014), 439-460

On the implication of structural zeros as independent variables in regression analysis: applications to alcohol research

Hua He[1*],Wenjuan Wang[1], Paul Crits-Christoph[2] , Robert Gallop[3], Wan Tang[1], Ding-Geng (Din) Chen[1] , Xin M. Tu[1]

[1]*University of Rochester Medical Center*
[2]*University of Pennsylvania*
[3]*West Chester University*

Abstract: In alcohol studies, drinking outcomes such as number of days of any alcohol drinking (DAD) over a period of time do not precisely capture the differences among subjects in a study population of interest. For example, the value of 0 on DAD could mean that the subject was continually abstinent from drinking such as lifetime abstainers or the subject was alcoholic, but happened not to use any alcohol during the period of interest. In statistics, zeros of the first kind are called structural zeros, to distinguish them from the sampling zeros of the second type. As the example indicates, the structural and sampling zeros represent two groups of subjects with quite different psychosocial outcomes. In the literature on alcohol use, although many recent studies have begun to explicitly account for the differences between the two types of zeros in modeling drinking variables as a response, none has acknowledged the implications of the different types of zeros when such modeling drinking variables are used as a predictor. This paper serves as the first attempt to tackle the latter issue and illustrate the importance of disentangling the structural and sampling zeros by using simulated as well as real study data.

Key words: number of days of drinking alcohol, NHANES, structural zero, zero-inflated count data, zero-inflated models for count data

1. Introduction

In alcohol studies or more generally in behavioral and psychosocial studies, it is important, both conceptually and methodologically, to pay special attention to structural zeros in count variables. Structural zeros refer to zero responses by those subjects whose count response will always be zero, in contrast to random (or sampling) zeros that occur to subjects whose count response can be greater than zero, but appear to be zero due to sampling variability. In alcohol research, count of days of alcohol use is commonly used to measure alcohol assumption.

* Corresponding author

Subjects who were always, or become, continually abstinent from drinking during a given time period form the non-risk subgroup of individuals with structural zeros in such drinking outcomes, while the remaining subjects constitute the at-risk subgroup. Such a partition of the study population is not only supported by the excess number of zeros observed in the distributions of drinking scores from many epidemiologic studies focusing on alcohol and related substance use (see Section 3.2), but also conceptually needed to serve as a basis for causal inference.

In the literature on alcohol use, this issue of structural zeros has been acknowledged when analyzing drinking variables as a response (Horton et al., 1999; Pardini et al., 2007). Although many studies, particularly older ones, failed to address the zero-inflated nature of their alcohol use variables, many more recent studies have done so across a wide range of alcohol use related variables (Buu et al., 2011; Connor et al., 2011; Cranford et al., 2010; Fernandez et al., 2011; Hagger-Johnson et al., 2011; Hernandez-Avila et al., 2006; Hildebrandt et al., 2010; Neal et al., 2005). All these articles focus on drinking outcomes when they are used as a response, using approaches such as the zero-inflated Poisson (ZIP) models (Hall and Zhang, 2004; Hall, 2000; Tang et al., 2012; Yu et al., 2012). The problem with structural zeros when such drinking measures are used as a predictor has not been addressed in the literature. The issue of structural zero in alcohol measures when used as predictor variables is equally, if not more, important to consider, especially when studying causal effects of drinking on health and behavioral outcomes. Usually count variables are treated as a continuous predictor, with no effort to distinguish structural zeros from their random counterparts. This practice is adopted for modeling convenience, which in many studies do not reflect the realistic association of variables involved. For example, a structure zero in a drinking outcome represents an individual who abstains from drinking, while a random zero corresponds to a drinker who did not drink during a period of time. Thus, the structural and random zeros represent two groups of subjects with quite different psychosocial outcomes. Beyond the field of alcohol research, another example of a predictor variable with structural zeros is the number of sexual partners in HIV/AIDS research, where structural zeros refer to those with lifetime celibacy or sexual problems, while random zeros are associated with those sexually active individuals who happened to have no sex during the study time.

One way to model the effect of a count variable with structure zeros as a predictor in regression analysis is to distinguish between random and structure zeros by including an indicator of structure zeros in the model, in addition to the count variable itself. This paper is aimed at illustrating the importance of this issue using simulation studies and how to deal with the issue using a real data example from the National Health and Nutrition Examination Survey (NHANES) database. The paper is organized as follows. In Section 2, we describe regression models with zero-inflated count variables as predictors and develop simulation studies to assess possible biases that may result when the effects of strucutral and random zero of a zero-inflated count predictor are not delineated. The results from simulation studies as well as a real data example are presented in Section 3, and the paper is concluded with a discussion in Section 4.

2. Materials and Methods

For notational brevity, we consider only the cross-sectional data setting. The considerations as well as the conclusions obtained apply to longitudinal study data as well. Given a sample of n subjects, let y_i denote the outcome of interest from the ith subject ($1 \le i \le n$). We are interested in assessing the effect of some personal trait such as alcohol dependency on the outcome, along with some other covariates, collectively denoted by $z_i = (z_{i1}, \ldots, z_{ip})^T$. Suppose that the trait is measured by a count variable χ_i with structure zeros.

Let r_i be the indicator of whether x_i is a structure zero, i.e., r_i=1 if x_i is a structure zero and r_i=0 otherwise. In some studies such as the real study considered in Section 3, the structure zeros are observed, which we assume throughout the paper unless stated otherwise. The indicator r_i partitions the study sample (population) into two distinctively different groups, with one consisting of subjects corresponding to r_i=1 and the other comprising of the remaining subjects with r_i =0. Since the trait in many studies is often a risk factor such as alcohol use, the first group is often referred to as the non-risk, while the second as the at-risk subgroup.

2.1 Linear and Generalized Linear Models

Without distinguishing between structural and random zeros, one may apply generalized linear models (GLM) to model the association between the explanatory variables, the predictor of interest x_i plus the covariates z_i, and the outcome. For example, if y_i is continuous, we may use the following linear model:

$$E(\ y_i \mid x_i, z_i\ ) = \alpha x_i + z_i^T \beta,\ 1 \le i \le n \tag{1}$$

Here one may include a covariate assuming a constant value 1 in z_i so that the intercept is included in β as well.

However, as mentioned in the Introduction, many count variables have structure zeros, which have quite a different conceptual interpretation than their random zero counterparts. This conceptual difference carries quite a significant implication for the interpretation of the coefficient α in (1). For example, within the context of drinking outcome, the difference in y_i between a subject with x_i = 1 and x_i = 0 has quite a different interpretation, depending on whether x_i = 0 is a structural or random zero. If x_i = 0 is a random zero, this difference represents the differential effect of drinking on y_i within the drinker subgroup when the drinking outcome changes from 0 to 1. For a structural zero, such a difference in y_i speaks to the effect of the trait of drinking on the response y_i. Thus, the model in (1) is flawed since it does not delineate the two types of effects and must be revised to incorporate the information of

structure zero. To this end, one may simply include an indicator of structural zero as a covariate in the model. By expanding (x_i, z_i) to include r_i, it follows from (1) that

$$E(\ y_i \mid x_i, z_i, r_i\ ) = \alpha x_i + z_i\beta + \gamma r_i, \quad 1 \le i \le n \tag{2}$$

Under the refined model above, the association between the trait and the response can be tested by checking whether both $\alpha = 0$ and $\gamma = 0$. This involves a composite linear contrast, H_0: $\alpha = 0,\ \gamma = 0$. If the null H_0 is rejected, then either $\alpha \neq 0$ or $\gamma \neq 0$ or $\alpha \neq 0$ or $\gamma \neq 0$. The coefficient γ is interpreted as the trait effect on the response y_i, all other things being equal. The coefficient α measures the change in y_i per unit increase in x_i within the at-risk group. For a binary or count response y_i, the linear model in (2) is readily extended using an appropriate member of the GLM. Instead of modeling $E(\ y_i \mid x_i, z_i, r_i\ )$ as a linear function of the explanatory variables, we assume that some function of $E(\ y_i \mid x_i, z_i, r_i\ )$, $h(E(\ y_i \mid x_i, z_i, r_i\ ))$, has a linear relationship with the explanatory variables. The choice of h, or link function, depends on whether y_i is a binary or count response. For example, the logit link is a popular choice for modeling a binary (logistic regression), while the log link is often used for a count y_i (Poisson log-linear model). The coefficients have the same conceptual interpretation, but their effects are interpreted in terms of odds ratio (logistic regression) and exponentiation (Poisson log-linear regression).

2.2 Zero-inflated Models

When the outcome y_i itself is a count response with structure zeros, it is not appropriate to apply Poisson or negative binomial (NB) log-linear models, the popular models for count responses. Instead, one needs to apply the zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB) model[Lambert, 1992; Tang et al., 2012; Welsh et al., 1996]. ZIP extends Poisson by including a logistic regression component so that it models both the at- and non-risk groups. Thus, estimates from the Poisson loglinear regression indicate increased/reduced effect of an explanatory variable on the count response of interest within the at-risk subgroup, while those from the logistic indicate increased/reduced risks of an explanatory variable for being in the non-risk subgroup, i.e., having the trait. By replacing Poison with NB, ZINB also addresses the weakness of the Poisson component in ZIP to account for overdispersion, a common violation of the Poisson that restricts the variance to be the same as the mean.

If ignoring the structure zero in x_i, one may model y_i using a ZIP:

$$\begin{aligned} &\text{structural zero } y_i \mid x_i, z_i \sim \text{Bernoulli}(\upsilon_i),\ \ \text{logit}\,(\upsilon_i) = \alpha' x_i + z_i^{\top}\beta', \\ &\text{non-structural zero count } y_i \mid x_i, z_i \sim \text{Poisson}(\mu_i),\ \ \log(\mu_i) = \alpha\, x_i + z_i^{\top}\beta, \end{aligned} \tag{3}$$

where Bernoulli(υ) (Poisson(μ)) denotes a Bernoulli (Poisson) distribution with mean υ (μ), and logit $(\upsilon) = \frac{\upsilon}{1+\upsilon}$ is the logit function. Under the ZIP above, the effect of x_i on the outcome

y_i is broken down into two parts, with one on the likelihood of being a structure zero, or being a member of the non-risk subgroup, determined by the logistic model in (3), and the other on the outcome y_i within the at-risk subgroup determined by the Poisson model in (3). Thus, one needs to test the null: H_0: $\alpha = \alpha' = 0$, to check if the trait is associated with the outcome y_i.

Similar to the linear model case, we can add as the indicator r_i of structure zeros of x_i as an additional predictor for the ZIP in (3) to obtain:

$$\text{structural zero } y_i \mid x_i, z_i, r_i \sim \text{Bernoulli}(v_i),\ \text{logit}(v_i) = \alpha' x_i + z_i^{\top}\beta' + \gamma' r_i,$$
$$\text{non-structural zero count } y_i \mid x_i, z_i, r_i \sim \text{Poisson}(\mu_i),\ \log(\mu_i) = \alpha x_i + z_i^{\top}\beta + \gamma r_I, \tag{4}$$

In this refined model, we need four coefficients to describe the relationship between the trait and the outcome. The coefficient γ measures the differential effect of the at- and non-risk group defined by x_i on the at-risk group defined by y_i, while the coefficient γ' captures the differential effect of the at- and non-risk group defined by x_i on the non-risk group defined by y_i. The coefficient α quantifies the increase in the outcome y_i within the at-risk group per unit increase in x_i within the at-risk subgroup defined by x_i, and the coefficient $\alpha'$is the log odds ratio for the change of likelihood of being in the non-risk group defined by *y* per unit increase in x_i among the at-risk subjects defined by x_i. If the trait and the outcome are not related, then all the four coefficients are zero:

$\alpha = \alpha' = \gamma = \gamma' = 0$. Note that for notational brevity, we have assumed no interaction among the explanatory variables in the models discussed above. In practice, some of these variables may also create interaction effect on the response y_i. Such interactions are readily incorporated into all the models discussed.

2.3 Simulation Studies

We performed simulation studies to show the importance of addressing structure zeros in studying the effect of a trait, measured by a count variable *X*, on a response of interest *Y*. The count predictor *X* was generated from a ZIP consisting of a Poisson with

mean μ and a point mass at zero, with a mixing probability of *p* (proportion of subjects with the trait. We considered four different types of resonse *Y*: continuous, binary, Poisson distributed count and zero inflated Poisson count response. We simulated data using the GLM for the first three cases. For the last case scenario, a Poisson variate was generated by GLM and then mixed with a constant zero based on the mixing probability of ZIP.

In all cases, the explanatory variables include *X* and the indicator *R* of structural zero of *X*. In addition to the true model, or Model I, which includes both *X* and *R*, we also considered Model II, which is identical to Model I, but with the indicator *R* removed. In the literature of alcohol studies, it is common to group drinking counts into categories before they are analyzed. Thus, we also dichotomized *X* according to whether *X* is positive. Thus, we also created Models III and IV by replacing *X* with such a dichotomized *X* in Models I and II, respectively.

A Monte Carlo (MC) size of 1,000 was used for all the models. We collected the point estimates of the coefficient of the count variable (*X*), and compared the bias and standard deviation of the estimates under the four different models. Further, we tested whether the trait was associated with the outcome *Y* and compared power across the models with type I error set at 0.05.

3. Results

3.1 Simulation Results

3.1.1 Continuous Response

For a continuous *Y*, the association of *Y* with *X* and *R* was specified as follows:

$$Y=c_0 + c_1X + c_2R + \varepsilon,\ \ \varepsilon\sim N(0,\sigma^2), \tag{5}$$

Where εis the error term. If c_1 and c_2 have different signs, say $c_1 > 0 > c_2$, then the mean of the at-risk subgroup defined by positive X > the mean of the non-risk group defined by structure zeros of *X*. In this case, this monotone relationships among the three subgroups will remain, even if the random and structure zeros are not distinguished between each other. However, if c_1 and c_2 have the same sign, say both are positive, $c_1, c_2 > 0$, then the mean of the at-risk subgroup defined by positive X > the mean of the at-risk group defined by random zeros of X < the mean of the non-risk group defined by structure zeros of *X*. In such cases, the mean of the non-risk group may be bigger than the at-risk subgroup defined by positive *X*, depending on the relationship between c_1 and c_2, and the monotone relationship among the three subgroups may fail, if random and structure zeros are combined. Thus, to assess power, we ran simulations to cover both situations, wherec_1 and c_2 had the same and different signs.

The zero inflated predictor *X* was simulated from a ZIP with the probability of being structural zero p = 0.2 and the mean of the Poisson component μ = 0.3. We simulated 1,000 samples with sample sizes of 100, 200, 500, and 1000, for several sets of parameters:

$$c_0 = 0.5, c_1 = -0.5, -0.25, 0, 0.25, 0.5,\quad c_2 = 0.5, \sigma^2 = 0.5 \tag{6}$$

For each simulated data, we fit the four aforementioned models,i.e.,

$$\text{Model I}\ :\ Y = c_0 + c_1X + c_2R + \varepsilon,\ \ \varepsilon\sim N(0,\sigma^2),$$

$$\text{Model II}\ :\ Y = c_0 + c_1X + \varepsilon,\ \ \varepsilon\sim N(0,\sigma^2)\ ,$$

$$\text{Model III}: Y = c_0 + c_1IX + c_2R + \varepsilon,\ \ \varepsilon\sim N(0,\sigma^2)\ ,$$

$$\text{Model IV}: Y = c_0 + c_1IX + \varepsilon,\ \ \varepsilon\sim N(0,\sigma^2),$$

where *I X* denotes the dichotomized *X* with *I X*=1(0) for $X > 0$ ($X \le 0$).

To save space, we only present some of the simulation results. Shown in Table 1 are the estimates (mean) of the parameters c_0, c_1, and c_2, and associated standard errors (Std err)

averaged over the 1,000 MC replications when sample size is 200. As expected the standard errors were similar between Models I and II as well as Models III and IV. However, the estimates from Model II (IV) were biased as compared to their counterparts from Model I (III). Although the estimates for parameters from Models I and III were not the same as their corresponding true values, the differences reflected the sampling variability. Note that the "true" value of the parameter c_1 under Model III should in fact be

$$E(\,Y \mid X > 0\,) - E(Y \mid X{=}0 \text{ and } R{=}0) = \frac{\mu c_1}{\Pr(X>0)} = \frac{0.3c_1}{1-\exp(-0.3)},$$

i.e., -0.58, -0.29, 0.00, 0.29, and 0.58, respectively, because of the grouping of subjects with $X > 0$.

Table 1: Parameter estimates (Mean) and standard errors (Std err) averaged over 1,000 MC replications for the four models considered in the Simulation Study with a continuous response.

Cases		c_0				c_1				c_2	
($c_1 =$)	Model	I	II	III	IV	I	II	III	IV	I	III
-0.5	Mean	0.50	0.62	0.50	0.63	-0.50	-0.60	-0.58	-0.70	0.49	0.49
	Std err	0.046	0.042	0.047	0.042	0.072	0.071	0.094	0.092	0.096	0.097
-0.25	Mean	0.50	0.62	0.50	0.63	-0.25	-0.35	-0.29	-0.42	0.49	0.49
	Std err	0.046	0.042	0.047	0.042	0.072	0.071	0.091	0.089	0.096	0.097
0	Mean	0.50	0.62	0.50	0.63	0.00	-0.10	0.00	-0.13	0.49	0.49
	Std err	0.046	0.042	0.047	0.042	0.072	0.071	0.090	0.088	0.096	0.097
0.25	Mean	0.50	0.62	0.50	0.63	0.25	0.15	0.29	0.16	0.49	0.49
	Std err	0.046	0.042	0.047	0.042	0.072	0.071	0.092	0.090	0.096	0.097
0.5	Mean	0.50	0.62	0.50	0.63	0.50	0.40	0.58	0.45	0.49	0.49
	Std err	0.046	0.042	0.047	0.042	0.072	0.071	0.096	0.094	0.096	0.097

Table 2: Estimated power in testing the association between the trait and the outcome based on 1,000 MC replications for the four models considered in the Simulation Study with a continuous response.

Cases ($c_1 =$)	Sample size	Model I	Model II	Model III	Model IV
−0.5	100	1.000	0.999	1.000	0.998
	200	1.000	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000
−0.25	100	0.997	0.879	0.997	0.872
	200	1.000	0.996	1.000	0.996
	500	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000
0	100	0.932	0.127	0.938	0.142
	200	0.999	0.227	0.999	0.249
	500	1.000	0.497	1.000	0.553
	1000	1.000	0.818	1.000	0.870
0.25	100	0.964	0.312	0.949	0.231
	200	0.999	0.547	0.998	0.407
	500	1.000	0.888	1.000	0.778
	1000	1.000	0.999	1.000	0.975
0.5	100	0.999	0.946	0.998	0.917
	200	1.000	0.999	1.000	0.999
	500	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000

Even if one does not care about the size of the effect of X on Y and just wants to detect association between the two variables, an application of the incorrect model such as Models II and Model IV may still be quite problematic. For example, we also examined power in detecting association between the trait and the outcome for the different models, with a type I error of 0.05. For Models II and IV, we can simply test the null: $H_0 : c_1 = 0$ for this purpose. However, for Models I and III, there are two terms pertaining to the association of interest, one relates to the difference between the structural and random zero in X (c_2) and the other is associated with difference between positive X and random zeros in X (c_1). So, we need to test a composite null: $H_0 : c_1 = c_2 = 0$ in Models I and III. We computed the proportions of p-values that were less than 0.05 for these hypothesis tests as the empirical power estimates. Shown in Table 2 are the estimated power to test the effect of the trait based on 1,000 MC replications with sample sizes 100, 200, 500 and 1000 in the range of values of c_1 (and c_2) considered. The models with the structure zero indicators included (Models I and III) were much more powerful in detecting the association between Y and X than their counterparts (Models II and IV). Thus, models that do not account for structural zeros such as Models II and IV may not even be able to perform such a "crude" task.

3.1.2 Binary Response *Y*

For a binary outcome Y , we simulated the response from a GLM with a logit link to relate X and R with Y :

$$Y \mid X, R \sim \text{Bernoulli}(v), \text{logit}(v) = c_0 + c_1 X + c_2 R. \quad (7)$$

Accordingly, we used four logistic models to fit the simulated data, akin to the four models in the continuous case, i.e., Model I and III included both X and R, while Model II and IV only included X. We again considered the four different sample sizes as in the continuous case, but report only the results for sample size = 200 for the parameter estimates for space consideration.

Table 3: Parameter estimates (Mean) and standard errors (Std err) averaged over 1,000 MC replications for the four models considered in the Simulation Study with a binary response.

Cases		c_0				c_1				c_2	
($c_1 =$)	Model	I	II	III	IV	I	II	III	IV	I	III
−0.5	Mean	0.51	0.62	0.51	0.62	−0.52	−0.60	−0.60	−0.71	0.49	0.49
	Std err	0.184	0.165	0.187	0.167	0.318	0.312	0.378	0.369	0.427	0.427
−0.25	Mean	0.51	0.61	0.51	0.62	−0.23	−0.32	−0.27	−0.38	0.49	0.49
	Std err	0.185	0.165	0.187	0.167	0.305	0.296	0.374	0.362	0.425	0.427
0	Mean	0.51	0.62	0.51	0.62	0.01	−0.08	0.01	−0.10	0.49	0.49
	Std err	0.184	0.166	0.187	0.167	0.315	0.307	0.380	0.368	0.423	0.427
0.25	Mean	0.51	0.62	0.51	0.62	0.28	0.19	0.32	0.21	0.49	0.49
	Std err	0.183	0.165	0.187	0.167	0.339	0.329	0.399	0.387	0.423	0.427
0.5	Mean	0.51	0.62	0.51	0.62	0.54	0.45	0.60	0.50	0.49	0.49
	Std err	0.184	0.166	0.187	0.167	0.369	0.367	0.422	0.413	0.424	0.427

Table 4: Estimated power in testing the association between the trait and the outcome based on 1,000 MC replications for the four models considered in the Simulation Study with a binary response.

Cases ($c_1 =$)	Sample size	Model I	Model II	Model III	Model IV
−0.5	100	0.301	0.263	0.291	0.272
	200	0.585	0.551	0.540	0.523
	500	0.936	0.891	0.912	0.860
	1000	0.997	0.996	0.996	0.990
−0.25	100	0.166	0.118	0.116	0.124
	200	0.275	0.191	0.258	0.192
	500	0.710	0.490	0.694	0.473
	1000	0.933	0.749	0.925	0.707
0	100	0.110	0.051	0.111	0.055
	200	0.163	0.060	0.154	0.055
	500	0.440	0.073	0.439	0.080
	1000	0.735	0.100	0.729	0.115
0.25	100	0.102	0.038	0.118	0.051
	200	0.190	0.062	0.185	0.069
	500	0.479	0.115	0.449	0.099
	1000	0.760	0.206	0.757	0.167
0.5	100	0.152	0.089	0.161	0.095
	200	0.316	0.205	0.300	0.193
	500	0.729	0.498	0.688	0.415
	1000	0.954	0.820	0.938	0.756

The averaged estimates of the parameters c_0, c_1, and c_2, and associated standard errors and power over the 1000 MC replications are shown in Tables 3 and 4. The standard errors were comparable between the corresponding models (Model I vs. II and Model III vs. IV), but the large bias observed in the continuous case persisted in each of the two incorrect models (Models II and IV). We again compared power between the correct (Models I and III) and incorrect (Models II and IV) models in detecting association between the trait and the response with the type I error set at 0.05. All power entries in Table 4 had much smaller values as compared the corresponding entries in Table 2, again confirming the fact that the incorrect models might incur a significant loss of power when detecting association between X and Y .

3.1.3 Poisson Count Response Y

For a Poisson distributed count variable Y, we generated Y from a GLM with a log function to link X and R to Y :

$$Y \sim \text{Poisson}\,(\mu),\ \log(\mu) = c_0 + c_1X + c_2R.$$

We fit four Poisson loglinear regression models to the data generated with the same set of parameter values as in the previous cases. We performed the simulation for each of the four

sample cases, but report the results for the case with sample size = 200 for the parameter estimates.

Table 5: Parameter estimates (Mean) and standard errors (Std err) averaged over 1,000 MC replications for the four models considered in the Simulation Study with a Poisson response.

Cases		c_0				c_1				c_2	
($c_1 =$)	Model	I	II	III	IV	I	II	III	IV	I	III
-0.5	Mean	0.50	0.65	0.50	0.65	-0.52	-0.65	-0.58	-0.74	0.51	0.51
	Std err	0.071	0.061	0.072	0.061	0.155	0.156	0.178	0.174	0.120	0.120
-0.25	Mean	0.50	0.65	0.50	0.65	-0.26	-0.38	-0.29	-0.45	0.51	0.51
	Std err	0.071	0.061	0.072	0.061	0.133	0.133	0.158	0.153	0.119	0.120
0	Mean	0.50	0.65	0.50	0.65	-0.01	-0.13	-0.01	-0.16	0.49	0.49
	Std err	0.070	0.060	0.072	0.061	0.114	0.114	0.142	0.137	0.119	0.120
0.25	Mean	0.50	0.65	0.50	0.65	0.25	0.13	0.29	0.13	0.49	0.49
	Std err	0.069	0.060	0.072	0.061	0.098	0.098	0.129	0.123	0.119	0.120
0.5	Mean	0.50	0.64	0.50	0.65	0.50	0.40	0.60	0.45	0.49	0.49
	Std err	0.068	0.060	0.071	0.061	0.082	0.082	0.123	0.116	0.119	0.120

Table 6: Estimated power in testing the association between the trait and the outcome based on 1,000 MC replications for the four models considered in the Simulation Study with a Poisson response.

Cases ($c_1 =$)	Sample size	Model I	Model II	Model III	Model IV
-0.5	100	0.970	0.901	0.967	0.885
	200	1.000	0.998	1.000	0.997
	500	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000
-0.25	100	0.901	0.533	0.897	0.531
	200	0.996	0.861	0.996	0.864
	500	1.000	0.999	1.000	0.999
	1000	1.000	1.000	1.000	1.000
0	100	0.752	0.090	0.755	0.109
	200	0.969	0.167	0.971	0.209
	500	1.000	0.395	1.000	0.433
	1000	1.000	0.675	1.000	0.733
0.25	100	0.797	0.218	0.769	0.161
	200	0.981	0.322	0.977	0.217
	500	1.000	0.651	1.000	0.471
	1000	1.000	0.900	1.000	0.746
0.5	100	0.976	0.869	0.958	0.809
	200	1.000	0.988	1.000	0.972
	500	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000

Shown in Table 5 are the averaged estimates of the parameters c_0, c_1, and c_2, and associated standard errors over the 1000 MC replications. The results show a similar pattern as in the previous cases; the standard errors were comparable between the corresponding models (Model

I vs. II and Model III vs. IV), but the large bias again remained in the two incorrect models (Models II and IV).

Shown in Table 6 are the estimated powers to test the effect of the trait based 1,000 MC replications the sample sizes 100, 200, 500 and 1000, in the range of values of c_1 (and c_2) considered. Similar to the previous cases, the power estimates also show a significant loss of power for the incorrect models (Models II and IV) when $-0.25 \leq c_1 \leq 0.25$. Note that although power was similar at c_1 =0.25 (−0.25), the settings were of no practical interest, since all power was close to 1. Note also that the power entries of Table 6 show that although not as powerful as the linear Y case, the Poisson Y was much more powerful than the binary Y to detect association between X and Y, all other things being equal.

3.1.4 Zero inflated Poisson Response Y

Finally, we considered a count response with structural zeros generated from the following ZIP:

$$\text{non-structural zero count } Y \mid X, R \sim \text{Poisson}(\mu), \log(\mu) = c_0 + c_1X + c_2R,$$
$$\text{structural zero } Y \mid X, R \sim \text{Bernoulli}(v), \text{logit}(v) = c_0 + c_1X + c_2R. \quad (8)$$

As in the previous cases, we fit four different ZIPs to the data simulated with the same set of parameter values (in addition to c_0, c_1 and c_2 are in previous cases, we set $c_0' = c_0, c_1' = c_1$ and $c_2' = c_2$). Again, we report the results for the case with sample size = 200 for the parameter estimates.

Table 7: Parameter estimates (Mean) and standard errors (Std err) averaged over 1,000 MC replications for the four models considered in the Simulation Study with a ZIP response.

Cases		c_0				c_1				c_2	
(c_1 =)	Model	I	II	III	IV	I	II	III	IV	I	III
−0.5	Mean	0.48	0.63	0.48	0.63	−0.55	−0.67	−0.60	−0.75	0.48	0.49
	Std err	0.155	0.126	0.155	0.126	0.370	0.360	0.425	0.416	0.298	0.299
−0.25	Mean	0.49	0.63	0.48	0.63	−0.30	−0.43	−0.33	−0.48	0.48	0.49
	Std err	0.152	0.125	0.155	0.126	0.330	0.325	0.392	0.383	0.297	0.299
0	Mean	0.49	0.63	0.48	0.63	−0.05	−0.17	−0.04	−0.19	0.48	0.49
	Std err	0.153	0.126	0.155	0.126	0.295	0.292	0.342	0.331	0.297	0.299
0.25	Mean	0.49	0.63	0.48	0.63	0.22	0.10	0.26	0.12	0.48	0.49
	Std err	0.150	0.125	0.155	0.126	0.266	0.263	0.315	0.302	0.295	0.299
0.5	Mean	0.49	0.63	0.48	0.63	0.49	0.37	0.57	0.42	0.48	0.48
	Std err	0.150	0.125	0.154	0.126	0.229	0.221	0.284	0.268	0.295	0.299

Table 8: Estimated power in testing the association between the trait and the outcome based on 1,000 MC replications for the four models considered in the Simulation Study with a ZIP response.

Cases ($c_1 =$)	Sample size	Model I	Model II	Model III	Model IV
−0.5	100	0.377	0.245	0.378	0.231
	200	0.611	0.393	0.567	0.301
	500	0.970	0.848	0.957	0.799
	1000	1.000	0.993	1.000	0.992
−0.25	100	0.276	0.115	0.274	0.112
	200	0.468	0.178	0.448	0.140
	500	0.903	0.468	0.900	0.439
	1000	0.996	0.813	0.996	0.794
0	100	0.219	0.078	0.207	0.058
	200	0.354	0.052	0.347	0.052
	500	0.763	0.055	0.759	0.074
	1000	0.969	0.136	0.969	0.154
0.25	100	0.213	0.090	0.201	0.081
	200	0.385	0.110	0.372	0.104
	500	0.783	0.190	0.748	0.117
	1000	0.983	0.359	0.971	0.240
0.5	100	0.352	0.283	0.309	0.243
	200	0.641	0.501	0.570	0.416
	500	0.975	0.851	0.946	0.774
	1000	1.000	0.993	0.999	0.966

Shown in Table 7 are the averaged estimates of the parameters c_0, c_1 and c_2, and associated standard errors over the 1000 MC replications. The same patterns of bias again emerged from the incorrect models (Models II and IV). The incorrect models also yielded much lower power than their correct counterparts. Shown in Table 8 are the estimated power to test the effect of the trait on the response. As seen, the ZIP seems to have similar power as the binary response Y, which is not surprising given that one of the components of ZIP is the binary response for modeling the structural zero of Y. Note that there are two components in ZIP models and thus the results are obtained from testing composite hypotheses. To see if the trait is associated with the outcome, we tested the null, $H_0 : c_1 = c_1' = c_2 = c_2' = 0$, for Models I and III.

3.2 A Case Study Example

We now illustrate the effect of bias in model estimates with a real data example based on the 2009-2010 National Health and Nutrition Examination Survey (NHANES). In this database, we identified a measure of alcohol use to be examined as an explanatory variable for depressive symptoms (count response). Both the alcohol and depression outcomes show a preponderance of zeros because of a large percent of the surveyed population is not at risk for either of the health issues. The relationship between the two has been reported in a number of studies

(Brown and Schuckit, 1988; Dackis et al., 1986; Davidson, 1995; Gibson and Becker, 1973; Hasin and Grant, 2002; Merikangas et al., 1998; Penick et al., 1988; Pettinati et al., 1982; Swendsen et al., 1998; Willenbring, 1986).

The NHANES is an annual national survey of the health and nutritional status of adults and children in the United States. A nationally representative sample of about 5,000 persons participates each year. Interviews and assessments are conducted in respondents' homes. Health assessments are performed in equipped mobile centers, which travel to locations throughout the country. Starting in 2007, NHANES has been oversampling all Hispanics (previously Mexican Americans were oversampled). In the 2009-2010 data set, data were collected from 10,537 individuals of all ages during the two-year period between January 2009 and December 2010. The race/ethnicity of the sample is 22.5% Hispanic-Mexican American, 10.8% Hispanic-other, 18.6% non-Hispanic Black, 42.1%, non-Hispanic White, and 6.1% other.

Alcohol Use Measure. In NHANES, for measurement of alcohol use, a different assessment was done for those aged 12 to 19 vs. those aged 20 and older; the assessment for the former age group asked only about the past 30 days, while the one administered to the latter age group asked about the past year. Therefore, for the current case study example we only used the data from the cohort aged 20 and older. Alcohol use (for those aged 20 or above) was assessed with a computer-assisted personal interview (CAPI). Specific questions of interest for the current work included number of days of any alcohol drinking (DAD) in the past year, which is commonly used in alcohol research. This variable was converted to average number of days drinking per month in our analysis. There were 6218 subjects in the data set with age of 20 and older.

In **NHANES**, one question asks "In your entire life, have you had at least 12 drinks of any type of alcoholic beverage?" This variable has been used previously to differentiate lifetime abstainers, who will answer "no" to this question and ex-drinkers, who will answer "yes", in NHANES (Tsai et al., 2012). Thus, the subjects who answered "no" to this question constitutes structural zeros. These results show that zero was endorsed by two distinctively different risk groups in this study population for the question about drinking.

Depression symptoms were measured in those aged 12 and above in the 2009-2010 NHANES with the Patient Health Questionnaire (PHQ-9) administered by CAPI. The PHQ-9 is a multiple-choice self-report inventory of 9 items specific to depression. Each item of the PHQ-9 evaluates the presence of one of the nine DSM-IV criteria for depressive disorder during the last two weeks. Each of the 9 items can be scored 0 (not at all), 1(few days), 2 (more than half the days) and 3 (nearly every day) and a total score is obtained. Among the 6218 subjects with CAPI, 5283 subjects reported PHQ-9, so there are about 935 subjects with missing values in the PHQ-9.

Covariates. In epidemiological samples, several demographic characteristics, including female gender, older age, not being married, low education, low income level, poor physical health, social isolation, minority status, and urban residence, have been associated with higher levels of depressive symptoms or presence of a major depressive disorder, though overlap among some of these factors suggests that these may not all be independent influences

(Gonz'alez et al., 2010; Leider-man et al., 2012; Oh et al., 2013; Roberts et al., 1997; Rushton et al., 2002; Weissman, 1996; Wilhelm et al., 2003). Based on these findings, in our analyses of the relationship of alcohol use to depressive symptoms, we incorporated relevant demographic variables available in NHANES (age, gender, education, race) as covariates.

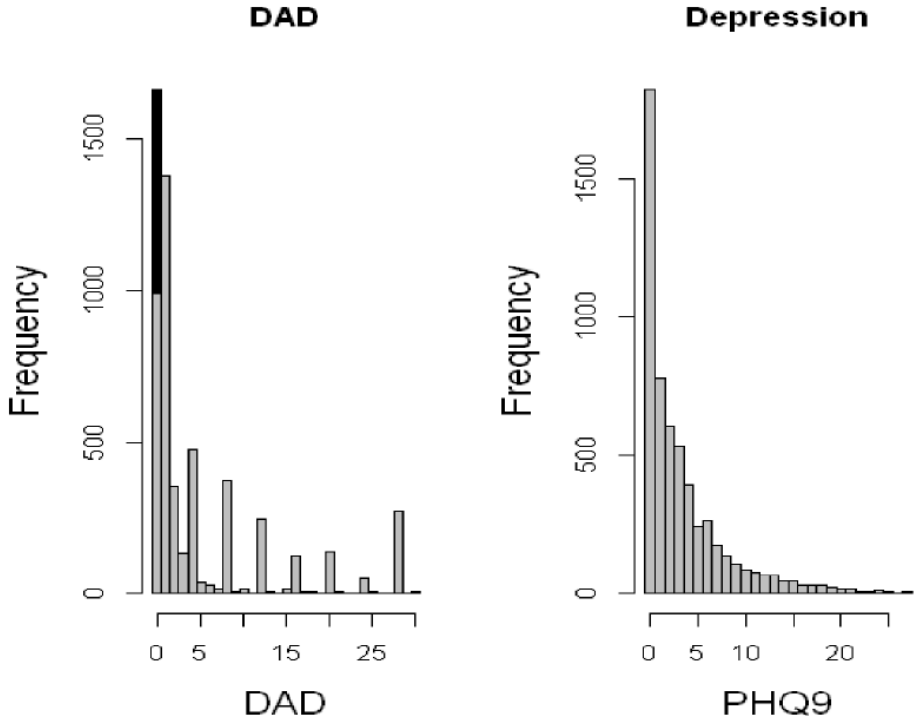

Figure 1: Distributions of DAD and PHQ9 for the 2009-2010 NHANES data, with the darker-shaded bar in the distribution of DAD representing structural zeros.

Shown in Figure 1 are the distributions of PHQ9 and DAD, both exhibiting a preponderance of zeros. Goodness of fit tests also rejected the fit of the data in each case by the Poisson (p-value < 0.001). Further, the Vuong test showed that ZIP provided a much better fit than the Poisson (p-value < 0.001). These findings are consistent with our prior knowledge that this study sample is from a mixed population consisting of an at-risk and non-risk subgroup for each of the behavioral and health outcomes.

Statistical Model. We applied the ZIP to model the PHQ-9 score with DAD in the past month as the predictor, adjusting for age, gender, race and education. Since we had the information to identify the non-risk group for the DAD variable, we conducted the analysis using two different models. In the first ZIP model, or Model I, we explicitly modeled the effect of structural zero of DAD on PHQ9 using a binary indicator (NeverDrink = 1 for structural and NeverDrink = 0 for sampling zero) and thus both the indicator of the non-risk group for drinking (NeverDrink) and DAD variable were included as predictors. As a comparison, we

also fit the data with only the DAD predictor and thus the structural and sampling zeros were not distinguished in the second ZIP, or Model II. We used SAS 9.3 PROC GENMOD to fit the models, with parameter estimates based on the maximum likelihood approach.

Analysis Result. Among the 5283 subjects with both CAPI and PHQ¬9, there were a small amount of missing values in the covariate and the actual sample size used for the analysis was 5261. Shown in Tables 9 and 10 are the parameter estimates of the Poisson and Zero Inflated components of the two ZIP models, respectively. The high statistical significance of the non-risk subgroup indicator in Model I indicates that Model I was more appropriate than Model II for the relationship of interest. In fact, Model I has a smaller AIC (28998.1975 for Model II vs. 28969.2253 for Model I) and BIC (29116.4228 for Model II vs. 29100.5868 for Model I).

Table 9: Comparison of model estimates (Estimate), standard errors (Std err) and p-values (P-value) fromthe Poisson component for the count response (including random zeros) (Std err) for the Real Study Data.

Parameter Estimates from Poisson Component of ZIP							
Parameter		Model I			Model II		
		Estimate	Std err	P-value	Estimate	Std err	P-value
Intercept		2.2936	0.0495	$< .0001$	2.2695	0.0490	$< .0001$
NeverDrink	Yes	-0.0878	0.0253	0.0005			
NeverDrink	No	0.0000	0.0000	.			
DAD		-0.0023	0.0011	0.0423	-0.0017	0.0011	0.1409
Gender	Male	-0.1988	0.0164	$< .0001$	-0.1890	0.0162	$< .0001$
	Female	0.0000	0.0000	.	0.0000	0.0000	.
AGE		-0.0025	0.0005	$< .0001$	-0.0026	0.0005	$< .0001$
Race/Ethnicity	Mexican American	-0.0954	0.0408	0.0192	-0.0928	0.0407	0.0228
	Other Hispanic	-0.0322	0.0425	0.4495	-0.0282	0.0425	0.5072
	Non-Hispanic White	-0.0952	0.0376	0.0114	-0.0864	0.0375	0.0213
	Non-Hispanic Black	0.0030	0.0401	0.9402	0.0087	0.0401	0.8278
	Other Race	0.0000	0.0000	.	0.0000	0.0000	.
Education		-0.1267	0.0067	$< .0001$	-0.1246	0.0067	$< .0001$
Scale		1.0000	0.0000		1.0000	0.0000	

Table 10: Comparison of model estimates (Estimate), standard errors(Std err) and p-values(P-value) from the Logistic component for the probability of occurrence of structural zeros for the Real Study Data.

Parameter Estimates from Logistic Component of ZIP							
Parameter		Model I			Model II		
		Estimate	Std err	P-value	Estimate	Std err	P-value
Intercept		−1.9351	0.1998	< .0001	−1.7999	0.1966	< .0001
NeverDrink	Yes	0.4230	0.0945	< .0001			
NeverDrink	No	0.0000	0.0000	.			
DAD		−0.0021	0.0043	0.6191	−0.0055	0.0042	0.1868
Gender	Male	0.5715	0.0641	< .0001	0.5174	0.0626	< .0001
	Female	0.0000	0.0000	.	0.0000	0.0000	.
AGE		0.0158	0.0018	< .0001	0.0164	0.0018	< .0001
Race/Ethnicity	Mexican American	0.0883	0.1596	0.5799	0.0594	0.1590	0.7087
	Other Hispanic	−0.0904	0.1694	0.5934	−0.1209	0.1688	0.4739
	Non-Hispanic White	−0.2250	0.1471	0.1262	−0.2714	0.1463	0.0637
	Non-Hispanic Black	0.0603	0.1563	0.6999	0.0309	0.1558	0.8428
	Other Race	0.0000	0.0000	.	0.0000	0.0000	.
Education		0.0480	0.0261	0.0658	0.0394	0.0259	0.1284

Based on the tables, the model without the indicator of the non-risk subgroup (Model II) failed to detect any association between DAD and depression symptoms (p-value=0.14 for the Poisson and 0.19 for the Zero-Inflated components), while the model with this indicator included (Model I) successfully identified a significant association between drinking and depression. The non-drinkers are less likely being at-risk of depression (p-value < 0.0001 for the Zero-Inflated component) and have less depressive symptoms (p-value=0.0005 for the Poisson component). On the other hand, DAD is also a predictor of depressive symptoms for the at-risk drinking subgroup (p-value = 0.0423 for the Poisson component). However, the amount of drinking does not seem to increase the likelihood of depression (p-value = 0.6191 for the Zero-Inflated component).

4. Discussion

In this paper, we discussed the importance of untangling the structural and random zeros in alcohol research. This is the first study to discuss this important issue in alcohol research. Attention to the zero-inflated nature of alcohol use measures has been mixed in the broad scientific literature. Although older studies completely ignored structural zeros, many newer ones have attempted to address this issue. However, all efforts to date have focused on the statistical problems when drinking outcomes are used as a response, or dependent variable, in regression analysis, with no attention paid to the equally important problem of biased estimates when such outcomes are used as an explanatory, or independent, variable. Indeed, this problem is not limited to alcohol research, since we failed to find any study in the extant literature that even acknowledged this problem. Our findings are significant in this respect since they show

for the time the critical importance of delineating the effects of the two different types of zeros in drinking outcomes like the DAD.

Both our simulated and real study examples demonstrate that it is critical that we model and delineate the effects of structural and random zero when using a zero-inflated count outcome as an explanatory variable in regression analysis. Otherwise, not only are we likely to miss opportunities to find association between such a variable and an outcome of interest (due to significant loss of power), but also to obtain results that are difficult to interpret because of high bias in the estimate and dual interpretion of the value zero of such a count variable. For example, the estimated coefficient −0.0017 of DAD in the Poisson component of ZIP Model II for the relationship between PHQ9 and DAD had about 30% upward bias as compared to −0.0023 for the same coefficient of the Poisson component of Model I ZIP of the analysis in the NHANES study. Even if ignoring such bias, the estimate −0.0017 was difficult to interpret; without accounting for structural zeros as in ZIP Model I, the change in DAD from 0 to 1 has a dubious meaning, since it may mean the change in amount of drinking within alcohol users or it may mean the difference between alcohol users vs. lifetime abstainers.

In all the examples considered, we assumed linear functions of explanatory variables for notational brevity. In practice, more complex functions of explanatory variables may be considered utilizing piecewise linear, polynomial functions or even nonparametric methods such as local polynomial regression. Also, we limited our considerations to cross-sectional studies, but the same considerations are readily applied to longitudinal studies.

We assumed that structural zeros of a count explanatory variable are known. In many studies, this may not be the case. For example, no lifetime abstinence was collected in NHANES for heavy drinking, another popular predictor for many behavioral and health outcomes. Thus, it is not possible to study the effect of the trait of heavy drinking on depression using the models considered in the study. Further research is needed to address this methodological issue to facilitate research in alcohol and other related areas.

Acknowledgements

The study was supported in part by National Institute on Drug Abuse grant R33DA027521, Eunice Kennedy Shriver National Institute of Child Health and Human Development grant R01HD075635, and a Faculty Research Support Grant from School of Nursing, University of Rochester.

References

[1] Brown, S. and Schuckit, M. (1988). Changes in depression among abstinent alcoholics. *Journal of Studies on Alcohol and Drugs*, **49**(05):412-417.

[2] Buu, A., Johnson, N., Li, R., and Tan, X. (2011). New variable selection meth¬ods for zero-in ated count data with applications to the substance abuse field. *Statistics in medicine*, **30**(18):2326-2340.

[3] Connor, J., Kypri, K., Bell, M., and Cousins, K. (2011). Alcohol outlet density, levels of drink-ing and alcohol-related harm in new zealand: a national study. *Journal of epidemiology and community health*, **65**(10):841-846.

[4] Cranford, J., Zucker, R., Jester, J., Puttler, L., and Fitzgerald, H. (2010). Parental alcohol in-volvement and adolescent alcohol expectancies predict alcohol involvement in male adolescents. *Psychology of Addictive Behaviors*, **24**(3):386-396.

[5] Dackis, C., Gold, M., Pottash, A., and Sweeney, D. (1986). Evaluating depression in alcoholics. *Psychiatry research*, **17**(2):105-109.

[6] Davidson, K. (1995). Diagnosis of depression in alcohol dependence: changes in prevalence with drinking status. *The British Journal of Psychiatry*, **166**(2):199¬204.

[7] 204. Fernandez, A., Wood, M., Laforge, R., and Black, J. (2011). Randomized trials of alcohol-use interventions with college students and their parents: lessons from the transitions project. *Clinical Trials*, **8**(2):205-213.

[8] Gibson, S. and Becker, J. (1973). Changes in alcoholics' self-reported depression. Quarterly *Journal of Studies on Alcohol*, 34:829-836.

[9] Gonz'alez, H. M., Tarraf, W., Whitfield, K. E., and Vega, W. A. (2010). The epidemiology of major depression and ethnicity in the united states. *Journal of psychiatric research*, **44**(15):1043-1051.

[10] Hagger-Johnson, G., Bewick, B., Conner, M., O'Connor, D., and Shickle, D. (2011). Alcohol, conscientiousness and event-level condom use. *British journal of health psychology*, **16**(4):828-845.

[11] Hall, D. and Zhang, Z. (2004). Marginal models for zero in ated clustered data. *Statistical Modelling*, **4**(3):161-180.

[12] Hall, D. B. (2000). Zero-in ated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**(4):1030-1039.

[13] Hasin, D. and Grant, D. (2002). Major depression in 6050 former drinkers -Asso-ciation with past alcohol dependence. *Archives of General Psychiatry*, **59**(9):794¬800.

[14] Hernandez-Avila, C., Song, C., Kuo, L., Tennen, H., Armeli, S., and Kranzler, H. (2006). Targeted versus daily naltrexone: secondary analysis of effects on average daily drinking. *Alcoholism: Clinical and Experimental Research*, **30(**5):860-865.

[15] Hildebrandt, T., McCrady, B., Epstein, E., Cook, S., and Jensen, N. (2010). When should clinicians switch treatments? an application of signal detection theory to two treatments for women with alcohol use disorders. *Behaviour research and therapy*, **48**(6):524-530.

[16] Horton, N., Bebchuk, J., Jones, C., Lipsitz, S., Catalano, P., Zahner, G., and Fitzmaurice, G. (1999). Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine*, **18**(2):213-222.

[17] Lambert, D. (1992). Zero-in ated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1):1-14.

[18] Leiderman, E. A., Lolich, M., V'azquez, G. H., and Baldessarini, R. J. (2012). Depression: Point-prevalence and sociodemographic correlates in a buenos aires community sample. *Journal of affective disorders*, **136**(3):1154-1158.

[19] Merikangas, K., Mehta, R., Molnar, B., Walters, E., Swendsen, J., Aguilar-Gaziola, S., Bijl, R., Borges, G., Caraveo-Anduaga, J., Dewit, D., et al. (1998). Comorbidity of substance use dis-orders with mood and anxiety disorders: results of the international consortium in psychiatric epidemiology. *Addictive Behaviors*, **23**(6):893-907.

[20] Neal, D., Sugarman, D., Hustad, J., Caska, C., and Carey, K. (2005). It's all fun and games... or is it? Collegiate sporting events and celebratory drinking. *Journal of Studies on Alcohol and Drugs*, **66**(2):291-294.

[21] Oh, D. H., Kim, S. A., Lee, H. Y., Seo, J. Y., Choi, B.-Y., and Nam, J. H. (2013). Prevalence and correlates of depressive symptoms in korean adults: Results of a 2009 korean community health survey. *Journal of Korean medical science*, **28**(1):128-135.

[22] Pardini, D., White, H., and Stouthamer-Loeber, M. (2007). Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood. *Drug and alcohol dependence*, **88**:S38-S49.

[23] Penick, E., Powell, B., Liskow, B., Jackson, J., and Nickel, E. (1988). The stability of coexisting psychiatric syndromes in alcoholic men after one year. *Journal of Studies on Alcohol and Drugs*, **49**(05):395-405.

[24] Pettinati, H., Sugerman, A., and Maurer, H. (1982). Four year mmpi changes in abstinent and drinking alcoholics. *Alcoholism: Clinical and Experimental Research*, **6**(4):487-494.

[25] Roberts, R. E., Kaplan, G. A., Shema, S. J., and Strawbridge, W. J. (1997). Does growing old increase the risk for depression? *American Journal of Psychiatry*, **154**(10):1384-1390.

[26] Rushton, J. L., Forcier, M., and Schectman, R. M. (2002). Epidemiology of depressive symptoms in the national longitudinal study of adolescent health. *Journal of the American Academy of Child Adolescent Psychiatry*, **41**(2):199-205.

[27] Swendsen, J., Merikangas, K., Canino, G., Kessler, R., Rubio-Stipec, M., and Angst, J. (1998). The comorbidity of alcoholism with anxiety and depressive dis-orders in four geographic communities. Comprehensive psychiatry, **39**(4):176-184.

[28] Tang, W., He, H., and Tu, X. (2012). Applied Categorical and Count Data Analysis. Chapman Hall/CRC, Boca Raton.

[29] Tsai, J., Ford, E., Li, C., and Zhao, G. (2012). Past and current alcohol con-sumption patterns and elevations in serum hepatic enzymes among us adults. *Addictive behaviors*, **37**:78-84.

[30] Weissman, M. e. a. (1996). Cross-national epidemiology of major depression and bipolar disorder. *Jama*, **276**(0):293-299.

[31] Welsh, A., Cunningham, R., Donnelly, C., and Lindenmayer, D. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**(1):297-308.

[32] Wilhelm, K., Mitchell, P., Slade, T., Brownhill, S., and Andrews, G. (2003). Prevalence and correlates of dsm-iv major depression in an australian national survey. *Journal of affective disorders*, **75**(2):155-162.

[33] Willenbring, M. (1986). Measurement of depression in alcoholics. *Journal of Studies on Alcohol*, **47**(5):367-372.

[34] Yu, Q., Chen, R., Tang, W., He, H., Gallop, R., Crits-Christoph, P., Hu, J., and Tu, X. (2012). Distribution-free models for longitudinal count responses with overdispersion and structural zeros. *Statistics in medicine*, **32**(14):2390-2405.

Received July 23, 2013; accepted February 23, 2014.

Hua He
University of Rochester Medical Center
265 Crittenden Boulevard, CU 420630, Rochester, New York 14642-0630
Hua_He@urmc.rochester.edu

Wenjuan Wang
University of Rochester Medical Center
601 Elmwood Avenue, P.O. Box 630, Rochester, NY 14642, USA

Paul Crits-Christoph
University of Pennsylvania
3535 Market St., Room 650
crits@mail.med.upenn.edu

Robert Gallop
West Chester University
25 University Avenue, West Chester, PA 19383
rgallop@wcupa.edu

Wan Tang
University of Rochester Medical Center
265 Crittenden Boulevard, CU 420630, Rochester, New York 14642-0630
Wan_Tang@urmc.rochester.edu

Ding-Geng (Din) Chen
University of Rochester Medical Center
265 Crittenden Boulevard, CU 420630, Rochester, New York 14642-0630
Din_Chen@urmc.rochester.edu

Xin M. Tu
University of Rochester Medical Center
265 Crittenden Boulevard, CU 420630, Rochester, New York 14642-0630
Xin_Tu@urmc.rochester.edu