Received: 22 July 2023 | Revised: 25 July 2024 | Accepted: 2 August 2024

DOI: 10.1002/sim.10198

RESEARCH ARTICLE

Statistics in Medicine WILEY

A simulation study of the performance of statistical models for count outcomes with excessive zeros

Zhengyang Zhou[1] | Dateng Li[2] | David Huh[3] | Minge Xie[4] | Eun-Young Mun[1]

[1]Department of Population and Community Health, University of North Texas Health Science Center, Fort Worth, Texas, USA

[2]Norden Lofts, White Plains, New York, USA

[3]School of Social Work, University of Washington, Seattle, Washington, USA

[4]Department of Statistics, Rutgers University, Piscataway, New Jersey, USA

Correspondence

Zhengyang Zhou, Department of Population and Community Health, University of North Texas Health Science Center, Fort Worth, TX, USA.
Email: zhengyang.zhou@unthsc.edu

Funding information

National Science Foundation, Grant/Award Numbers: DMS2015373, DMS2027855, DMS2311064, DMS2319260; National Institute on Alcohol Abuse and Alcoholism, Grant/Award Numbers: K02 AA028630, R01 AA019511; National Institute of Allergy and Infectious Diseases, Grant/Award Number: R21 AI179404

Background: Outcome measures that are count variables with excessive zeros are common in health behaviors research. Examples include the number of standard drinks consumed or alcohol-related problems experienced over time. There is a lack of empirical data about the relative performance of prevailing statistical models for assessing the efficacy of interventions when outcomes are zero-inflated, particularly compared with recently developed marginalized count regression approaches for such data. **Methods**: The current simulation study examined five commonly used approaches for analyzing count outcomes, including two linear models (with outcomes on raw and log-transformed scales, respectively) and three prevailing count distribution-based models (ie, Poisson, negative binomial, and zero-inflated Poisson (ZIP) models). We also considered the marginalized zero-inflated Poisson (MZIP) model, a novel alternative that estimates the overall effects on the population mean while adjusting for zero-inflation. Motivated by alcohol misuse prevention trials, extensive simulations were conducted to evaluate and compare the statistical power and Type I error rate of the statistical models and approaches across data conditions that varied in sample size ($N = 100$ to 500), zero rate (0.2 to 0.8), and intervention effect sizes. **Results**: Under zero-inflation, the Poisson model failed to control the Type I error rate, resulting in higher than expected false positive results. When the intervention effects on the zero (vs. non-zero) and count parts were in the same direction, the MZIP model had the highest statistical power, followed by the linear model with outcomes on the raw scale, negative binomial model, and ZIP model. The performance of the linear model with a log-transformed outcome variable was unsatisfactory. **Conclusions**: The MZIP model demonstrated better statistical properties in detecting true intervention effects and controlling false positive results for zero-inflated count outcomes. This MZIP model may serve as an appealing analytical approach to evaluating overall intervention effects in studies with count outcomes marked by excessive zeros.

KEYWORDS

count outcome, marginalized model, simulation, statistical power, type I error, zero inflation

1 | INTRODUCTION

Count outcomes are frequently encountered in health behaviors research. Examples of such data include number of standard drinks containing alcohol consumed,[1] number of cigarettes smoked,[2] and number of sexual risk behaviors experienced.[3] Zero-inflation occurs when there is an excessive proportion of outcome values stacked at zero, which is a common phenomenon, especially with behavioral health outcomes. For example, in alcohol prevention and intervention trials aimed at reducing alcohol consumption among participants, the proportion of participants reporting zero alcohol drink can be as high as 66%, suggesting that the outcome variable was zero inflated.[4] One reason for the disproportionate proportions of zeros and non-zero values is that the study population may consist of two clinically distinct groups, where one group comprises participants at-risk for alcohol consumption and the other comprises participants not-at-risk (eg, participants who are abstainers that will not consume any drink, resulting in zero-inflation). In the following context, we refer to the above two groups as the *at-risk* and *not-at-risk* subpopulations, respectively.

Many types of statistical approaches have been utilized to model count data in the literature. To appropriately account for the count nature of such data, researchers have used generalized linear models based on count distributions, such as the Poisson and negative binomial (NB). The Poisson regression model assumes that the mean of the outcome is equal to the variance, while the NB regression model allows the variance to be greater than the mean by incorporating an additional dispersion parameter. Both the Poisson and NB models assume that the study sample comes from one homogeneous population and relate covariates to the mean outcome of the entire sample. With these models, there is no flexibility to account for excessive zeros when the count outcome of interest is zero inflated.

The zero-inflated Poisson (ZIP) model is an extension of the regular Poisson model that is more appropriate for count data with excessive zeros by using a mixture distribution of the Poisson distribution and a point mass at zero (ie, the structural zeros). In the context of alcohol intervention trials, the Poisson part can be considered as evaluating the *at-risk* subpopulation who may or may not drink at a given assessment, and the structural zero part as evaluating the *not-at-risk* subpopulation who "predictably" do not drink (eg, abstainers for religious or other personal reasons). More discussion for the two-part nature of the ZIP model can be found in Reference 5. Unlike the Poisson and NB models that evaluate the effects of each covariate on the overall mean of the outcome, the ZIP model separately evaluates the effects on the two parameters of the mixture distribution–the Poisson mean and the probability of a structural zero. As a consequence, the estimates from a ZIP model can be cumbersome to interpret, as they describe two different parameters for two subpopulations.

To directly infer the effects on the overall mean and maintain the ability to account for zero inflation, the marginalized ZIP (MZIP) model has been proposed based on the framework of the ZIP model.[6-8] Instead of separately evaluating the effects on the two parameters, this approach makes direct inference on the overall effect of the entire sample by linking the marginal (or overall) mean of the outcome to the covariates. Compared to the ZIP model, which conceptually separates the population into two subpopulations, the MZIP model treats the entire sample as a whole, which makes it feasible to answer the following, simpler but often critical question of *whether the intervention is efficacious for the entire study population*. That question is commonly of principal interest when a clinical trial is designed[9] and can be accounted for in a calculation of sample size based on an MZIP model.[10]

Despite the increasing availability of new statistical methods and software for analyzing count data with zero inflation, a nonignorable number of studies still do not utilize appropriate statistical methods for such data. For example, in a meta-analysis of 17 studies using individual participant data from each, over half (nine studies) had excessive proportions of zero outcomes (ie, number of drinks). However, of the nine, eight did not account for this zero inflation.[4] A review by Reference 11 summarized the statistical models used to evaluate the effectiveness of brief alcohol interventions in reducing alcohol consumption. The investigators reviewed 119 alcohol-related count outcomes from 64 papers and observed that more than half of the outcomes (61.3%) were analyzed using statistical models that assume normally distributed residuals. Less than a third (31.1%) were analyzed using count distribution models. These observations suggest a gap between the methodological advances and their applications in applied research.

In this article, we aim to bridge the implementation gap by providing evidence-based guidance in selecting appropriate statistical models for count data with or without excessive zeroes through extensive simulation studies. To this end, we conducted a methodological phase III study[12] to compare the statistical performance across candidate methods for count data by evaluating their ability to control Type I error (ie, ability to control false positive findings) and statistical power (ie, ability to detect a true effect, when it exists) under a large range of data conditions inferred from application settings. Based on the empirical evidence, we then provide pragmatic recommendations on selecting a preferred method among candidate methods in different scenarios (eg, when a certain method is preferred). More specifically, we consider three broad sets of statistical models. The first set consists of conventional count distribution-based models for data with or without zero-inflation, including the Poisson, NB, and ZIP models. The second is a marginalized model for zero-inflated data, specifically the MZIP model. The third set consists of linear models, with or without logarithm transformation, which have been commonly used in the literature (eg, see Reference 11). This current study is a neutral comparison study, where we focus on comparing existing approaches for count data, without preference for any particular method. None of the authors of the current study was involved in the development of the methods being evaluated. Note, however, that our group developed an R package "mcount"[13] to fit the MZIP model based on methods as described in the original paper[7] for applications. For more discussion on the concept of neutral comparison study, please see References 14,15.

This article is organized as follows. In Section 2, we describe the formulation of the count distribution-based models considered in the simulation study. In Section 3, we describe a simulation study to evaluate and compare the relative performances of candidate methods under various data conditions. In Section 4, we summarize the empirical results in terms of Type I error and statistical power obtained from the simulation study. In Section 5, we discuss the overall findings and conclusions.

2 | METHODS

For a clinical trial with two arms, let us assume that the outcome of interest is a count variable that may or may not have excessive zero values. Suppose that the study sample size is n and for the i-th participant, $i = 1, 2, \dots, n$, the count outcome is y_i. Consider $p - 1$ covariates in the statistical model of the trial outcome, one of which is the intervention assignment indicator $\mathbb{1}_{\{A_i=T\}}$, where A_i denotes i-th participant's assignment to either the intervention (T) or control (C) arm. Denote the remaining $p - 2$ covariates as $\mathbf{x}_{i,p-2} = (x_{i2}, x_{i3}, \dots, x_{i,p-1})^t$. In the following, we describe potential statistical models that may be considered to evaluate the intervention effect on the count outcome, including the Poisson, NB, ZIP, MZIP, and linear regression models with raw scale scores or log-transformed scores.

2.1 | Poisson and NB regression models

Among the count distribution-based regression models, the Poisson model has the most straightforward formulation by modeling the logarithm of the mean outcome through a list of predictors. The outcome values are assumed to follow the Poisson distribution, which restricts the mean value to be equal to its variance. When there is "overdispersion" in the data, where the variance of the distribution is larger than the mean, the Poisson model can underestimate variance and yield invalid inferences. Based on the two-arm trial design described at the beginning of the Methods Section, the Poisson model can be expressed as

$$\begin{aligned} &y_i \sim \text{Poisson}(\nu_i), \\ &\log(\nu_i) = \mathbf{x}_i^t \boldsymbol{\beta}^{Poi} = \beta_0^{Poi} + \beta_1^{Poi} \mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\eta}^{Poi}, \end{aligned} \tag{1}$$

where $\nu_i = \mathbb{E}[y_i]$ is the overall mean of the outcome under a Poisson distribution, $\mathbf{x}_i = (1, \mathbb{1}_{\{A_i=T\}}, \mathbf{x}_{i,p-2}^t)^t$, and $\boldsymbol{\beta}^{Poi} = (\beta_0^{Poi}, \beta_1^{Poi}, \boldsymbol{\eta}^{Poi(t)})^t = (\beta_0^{Poi}, \beta_1^{Poi}, \beta_2^{Poi}, \dots, \beta_{p-1}^{Poi})^t$ are the vectors of regressors and regression coefficients, respectively.

The NB model is an alternative count regression model. Compared to the Poisson regression model, it incorporates an additional "dispersion" parameter, which allows the variance to be greater than the mean. Therefore, the NB regression model is flexible in accommodating overdispersion. Similarly, the NB regression model can be expressed as

$$\begin{aligned} &y_i \sim \text{NB}(\nu_i, k), \\ &\log(\nu_i) = \mathbf{x}_i^t \boldsymbol{\beta}^{NB} = \beta_0^{NB} + \beta_1^{NB} \mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\eta}^{NB}, \end{aligned} \tag{2}$$

where $\nu_i = \mathbb{E}[y_i]$ is the overall mean of the outcome and $k > 0$ is the dispersion parameter, which satisfies $\text{Var}[y_i] = \nu_i + \frac{\nu_i^2}{k}$. Of note, the above parameterization for dispersion follows the Type 2 NB distribution (or NB2) with a quadratic mean-variance relationship.[16] For a Type 1 NB distribution parameterization, the variance is a linear function of the mean, which is less commonly used in practice.

2.2 ZIP regression model

When evaluating count outcomes, zero inflation is said to be present when the observed proportion of zeros is much greater than the theoretically expected proportion under a conventional count distribution, such as Poisson or NB. Such an observation typically reflects inherent study population characteristics, where a subset of participants will “predictably” produce zeros (ie, the *not-at-risk* subpopulation as previously described in the Introduction section). In the presence of zero inflation, the Poisson and NB models may not perform well because their formulations do not account for excessive zeros. As a result, the two statistical models (ie, Poisson and NB models) could produce biased effect size estimates and inaccurate statistical significance inferences.[17]

The ZIP model was proposed to account for zero inflation by explicitly modeling excessive zeros.[18,19] This model assumes that a count outcome follows a mixture distribution consisting of the Poisson distribution and a point mass at zero (ie, the structural zeros). For example, in the context of alcohol intervention trials whose primary goal is to reduce the number of drinks consumed, the Poisson part can be considered as evaluating an *at-risk* subpopulation who may or may not drink at a given assessment. In contrast, the structural zero part can be considered as evaluating the *not-at-risk* subpopulation who “predictably” do not drink (eg, abstainers for religious or other personal reasons). More discussion on the two-part nature of the ZIP model can be found in Reference 5. Consequently, the intervention effects are estimated in two separate parts–the rate ratio (RR) of the mean in the Poisson part (eg, number of drinks, including random zeros from those who happened not to drink) and the odds ratio (OR) of being a structural zero (eg, abstainers vs. non-abstainers) in the structural zero part. The ZIP model can be formally expressed as

$$
\begin{aligned}
&y_i \sim \begin{cases} 0 & \text{with probability } \pi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1-\pi_i \end{cases}, \\
&\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^t \boldsymbol{\gamma}^{ZIP} = \gamma_0^{ZIP} + \gamma_1^{ZIP}\mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\zeta}^{ZIP}, \text{ and} \\
&\log(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta}^{ZIP} = \beta_0^{ZIP} + \beta_1^{ZIP}\mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\eta}^{ZIP},
\end{aligned} \tag{3}
$$

where $\mu_i = \mathbb{E}[y_i | y_i \text{ from the Poisson part}]$ is the mean parameter of the Poisson part, $\pi_i = Pr(y_i \text{ is a structural zero})$ is the structural zero rate, $\boldsymbol{\beta}^{ZIP} = (\beta_0^{ZIP}, \beta_1^{ZIP}, \boldsymbol{\eta}^{ZIP(t)})^t = (\beta_0^{ZIP}, \beta_1^{ZIP}, \beta_2^{ZIP}, \dots, \beta_{p-1}^{ZIP})^t$ are the regression coefficients for the Poisson part, and $\boldsymbol{\gamma}^{ZIP} = (\gamma_0^{ZIP}, \gamma_1^{ZIP}, \boldsymbol{\zeta}^{ZIP(t)})^t = (\gamma_0^{ZIP}, \gamma_1^{ZIP}, \gamma_2^{ZIP}, \dots, \gamma_{p-1}^{ZIP})^t$ are the regression coefficients for the structural zero part. When applying the ZIP model to data, covariate effects, such as an intervention or treatment effect, are interpreted separately for the two parts, corresponding to two distinct subpopulations. However, in many clinical trials, whether there is an “overall” intervention or treatment effect for the entire population is important and often the principal clinical question. Unfortunately, the overall intervention effect is not straightforward to evaluate in a ZIP model. Of note, under the ZIP model, the “overall mean,” which is denoted by $\mathbb{E}[y_i] \triangleq \nu_i$, can be expressed as

$$
\nu_i = (1-\pi_i)\mu_i = \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}^{ZIP}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\gamma}^{ZIP}}}. \tag{4}
$$

Equation (4) implies that the “overall mean” depends on all covariates and consequently all parameters from the two parts of the model. More importantly, the overall effect of the intervention, which is usually defined as the incidence rate ratio (IRR) between the intervention (T) and control (C) groups holding other covariates constant, is expressed as

$$
\frac{\mathbb{E}[y_i | A_i = T, \mathbf{x}_{p-2}]}{\mathbb{E}[y_j | A_j = C, \mathbf{x}_{p-2}]} = \exp(\beta_1^{ZIP}) \frac{1 + \exp(\gamma_0^{ZIP} + \mathbf{x}_{p-2}^t \boldsymbol{\zeta}^{ZIP})}{1 + \exp(\gamma_0^{ZIP} + \gamma_1^{ZIP} + \mathbf{x}_{p-2}^t \boldsymbol{\zeta}^{ZIP})}, \tag{5}
$$

where i and j represent two hypothetical participants in treatment and control groups, respectively, and with the same sets of covariates (ie, $\mathbf{x}_{p-2}$). Because the IRR is a function of covariates $\mathbf{x}_{p-2}$ in Equation (5), it implies that under the ZIP model, unless the treatment indicator is the only covariate included in the model (ie, $\mathbf{x}_{p-2} = 0$), the intervention effect depends on all other covariate values and varies across individuals. To obtain a population-level overall treatment effect from the ZIP model, it is necessary to integrate out all covariates, which can be computationally tedious and error prone.

2.3 | MZIP model

The MZIP model is an extension of the ZIP model.[6-8] The MZIP model accounts for zero inflation and directly models the overall mean of the outcome. Recall that we denote $\mathbb{E}[y_i] \triangleq \nu_i$ as the overall mean and μ_i as the mean of the Poisson variable. The MZIP model is then expressed as

$$\begin{aligned}
&y_i \sim \begin{cases} 0 & \text{with probability } \pi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1-\pi_i \end{cases}, \\
&\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP} = \gamma_0^{MZIP} + \gamma_1^{MZIP} \mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\zeta}^{MZIP}, \text{ and} \\
&\log(\nu_i) = \mathbf{x}_i^t \boldsymbol{\beta}^{MZIP} = \beta_0^{MZIP} + \beta_1^{MZIP} \mathbb{1}_{\{A_i=T\}} + \mathbf{x}_{i,p-2}^t \boldsymbol{\eta}^{MZIP}.
\end{aligned} \tag{6}$$

Note that the MZIP model is different from the ZIP model in that the MZIP *directly* models the overall mean (ie, ν_i) through covariates, instead of the mean in the Poisson part (ie, μ_i) as in the ZIP model (see Equations 4 and 5). No additional regression equation is needed for the Poisson mean, μ_i, as it is determined through the equation $\mu_i = \frac{\nu_i}{1-\pi_i}$. Intuitively, to produce the same overall mean of the entire population (ν_i), the mean number of outcomes in the *at-risk* subpopulation (μ_i) would need to be higher to offset a scenario with a greater proportion of individuals in the *not-at-risk* subpopulation (π_i), and vice versa.

In Equation (6), the intervention effect on the “overall mean” outcome for the entire population is quantified by β_1^{MZIP}, which can be interpreted as the log incidence density ratio difference between the intervention and control groups. Therefore, β_1^{MZIP} enjoys the same straightforward interpretation as β_1^{Poi} or β_1^{NB}. Compared to the ZIP model where the intervention effect is interpreted separately for the *at-risk* (through β_1^{ZIP}) and *not-at-risk* (through γ_1^{ZIP}) subpopulations, the MZIP model evaluates the intervention effect on the entire population through the latent parameter β_1^{MZIP}. With a single intervention effect estimate on the entire population, the MZIP model is advantageous over the ZIP model when answering the question of whether, and to what extent, an intervention in question is efficacious for the entire population. In addition to the intervention effect on the overall mean estimate, the MZIP model provides the parameter estimates in the structural zero part using the same formula as in the ZIP model. Therefore, the MZIP model retains the ability to evaluate the intervention effect for the *not-at-risk* subpopulation through the latent parameter γ_1^{MZIP}.

The likelihood function of an MZIP model is as follows:

$$\begin{aligned}
L(\boldsymbol{\gamma}^{MZIP}, \boldsymbol{\beta}^{MZIP} | \boldsymbol{y}) = &\prod_{y_i} (1 + e^{\mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP}})^{-1} \prod_{y_i=0} [e^{\mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP}} + e^{-(1+\exp(\mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP}))\exp(\mathbf{x}_i^t \boldsymbol{\beta}^{MZIP})}] \\
&\prod_{y_i>0} \frac{[e^{-(1+\exp(\mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP}))\exp(\mathbf{x}_i^t \boldsymbol{\beta}^{MZIP})} (1 + e^{\mathbf{x}_i^t \boldsymbol{\gamma}^{MZIP}})^{y_i} e^{y_i \mathbf{x}_i^t \boldsymbol{\beta}^{MZIP}}]}{y_i!}
\end{aligned} \tag{7}$$

To fit an MZIP model, one can estimate the parameters by maximizing the likelihood function shown in Equation (7) using non-linear optimization algorithms, which has been implemented in an R package “mcount” (Reference 13; see Reference 9 for a real data application utilizing the “mcount” R package).

3 | SIMULATION

The simulation study is structured according to the ADEMP scheme, which has five elements, including aims, data-generating mechanisms, estimands and other targets, methods, and performance measures.[20] Each of the five elements is described in the following subsections.

3.1 | Aims

The simulation study is aimed at evaluating comparative performances across five statistical models for count data in terms of empirical statistical power and Type I error in various data situations. The methods considered are the Poisson model, NB model, ZIP model, MZIP model, and linear model.

3.2 | Data-generating mechanisms

The simulation settings for study characteristics were based on motivating data from Project INTEGRATE,[21] a large-scale individual participant data meta-analysis project examining the effectiveness of brief alcohol interventions on reducing alcohol consumption among young adults. Therefore, the simulation settings used in this study represent a broad range of data conditions in this field. Because the ZIP regression model allows for the flexible manipulation of the intervention effect on the two subpopulations through the Poisson and structural zero parts, it was selected as the data generating model. Specifically, for an individual study with two arms, we considered total sample sizes of $N \in \{100{,}200, 300{,}500\}$, where the outcome of the i-th subject ($i \in \{1, 2, \ldots, N\}$) was simulated by a ZIP model characterized by $y_i \sim \text{Poisson}(\mu_i)$ with probability $1 - \pi_i$, and 0 otherwise. The Poisson mean parameter μ_i and the structural zero rate π_i were determined through the following link functions

$$\begin{aligned} &\log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_0 + \gamma_1 \mathbb{1}_{\{A_i=T\}} + \gamma_2 \text{Cov}_i, \text{ and} \\ &\log(\mu_i) = \beta_0 + \beta_1 \mathbb{1}_{\{A_i=T\}} + \beta_2 \text{Cov}_i, \end{aligned} \tag{8}$$

where the intervention group assignment was determined by $\mathbb{1}_{\{A_i=T\}} \sim \text{Bernoulli}(0.5)$ and the covariate was generated by $\text{Cov}_i \sim N(0, 1)$. Equation (8) is a special case of the general formulation of a ZIP model (Equation 3). For this simulation, we considered the situation, in which the outcomes are explained by the difference in the intervention versus control group and an additional individual-level covariate for baseline differences in the outcome.

The regression coefficients β_1 and γ_1 in Equation (8) measure the intervention effects on the Poisson and structural zero parts of the ZIP model, respectively. In the context of alcohol intervention studies, β_1 could quantify the effect of the intervention on the average number of drinks for participants who drink, with $\beta_1 < 0$ representing a favorable intervention effect (ie, reduced drinking). γ_1 would quantify the intervention effect on the proportion of abstainers, with $\gamma_1 > 0$ indicating a favorable intervention effect (ie, a greater proportion of non-drinking) in the intervention arm. Since the intervention influences the outcome in two ways, we considered the following four intervention conditions characterized by different values of β_1 and γ_1:

Condition 1. $\beta_1 \in \{-0.1, -0.2, -0.3\}$ and $\gamma_1 = 0.5$: Intervention *reduces* the average number of drinks for those who drink (ie, RR = 0.90, 0.82, 0.74) and *increases* the proportion of abstainers or nondrinkers (ie, OR = 1.65).

Condition 2. $\beta_1 \in \{-0.1, -0.2, -0.3\}$ and $\gamma_1 = 0$: Intervention *reduces* the average number of drinks for those who drink (ie, RR = 0.90, 0.82, 0.74) but *has no effect on* abstainers (ie, OR = 1).

Condition 3. $\beta_1 = 0$ and $\gamma_1 = 0.5$: Intervention *has no effect on* the average number of drinks (ie, RR = 1) for those who drink but *increases* the proportion of abstainers (ie, OR = 1.65).

Condition 4. $\beta_1 \in \{0.1, 0.2, 0.3\}$ and $\gamma_1 = 0.5$: Intervention *increases* the average number of drinks for those who drink (ie, RR = 1.11, 1.22, 1.35) and *increases* the proportion of abstainers or nondrinkers (ie, OR = 1.65).

Condition 5. $\beta_1 = 0$ and $\gamma_1 = 0$: Intervention *has no effect on* both the average number of drinks and the proportion of abstainers (ie, RR = 1 and OR = 1).

Conditions 1–4 will be used to evaluate statistical power across the models, where intervention is effective for the number of drinks, the likelihood of drinking, or both. In Condition 4, the intervention has an iatrogenic effect increasing the average number of drinks consumed ($\beta_1 > 0$) but has an intended effect increasing the proportion of non-drinking ($\gamma_1 > 0$). Because the intervention has conflicting effects on the two parts, this condition may be less common in alcohol prevention and intervention trials. However, this atypical scenario provides interesting empirical evidence on the relative performance of the methods, which we will discuss later. In Condition 5, the intervention has no effect. This condition can be used to evaluate the Type I error rate of all models considered.

To evaluate statistical properties over different levels of zero inflation, we varied the proportion of zero outcomes from 0.2, 0.3, …, 0.8. Once β_1 and γ_1 are determined in each condition, we set $\beta_0 = 0.8 - \beta_1$ and $\beta_2 = 0.2$, so that the mean expected number of drinks is fixed, ensuring that samples are comparable across simulation conditions with respect to their drinking level. We further constrained $\gamma_0 = 2\gamma_2$, so γ_2 can be calculated such that the pre-determined zero rate can be reached. Note that β_2 and γ_2 are the parameters quantifying the effects of a covariate in the simulation, which are of less interest in applied research. The simulation settings are summarized in Table 1.

TABLE 1 Summary of simulation settings.

	Possible values
Sample size	100, 200, 300, 500
Zero rate	0.2, 0.3, …, 0.8
β_1	−0.1, −0.2, −0.3, 0, 0.1, 0.2, 0.3
γ_1	0, 0.5

3.3 | Estimands and other targets

The simulation estimand or target of interest is to test a null hypothesis regarding the intervention effect, that is, H_0: *the intervention has no effect on alcohol consumption*. This is a general expression and the specific null hypothesis for each method is described below in Section 3.4.

3.4 | Methods

We considered the Poisson, NB, ZIP, MZIP, and linear models in this simulation. The null hypothesis to test intervention effects in each method was described as follows.

1. Poisson model–testing the effect of intervention on the overall mean of the entire population ($H_0 : \beta_1^{Poi} = 0$).
2. NB model–testing the effect of intervention on the overall mean of the entire population ($H_0 : \beta_1^{NB} = 0$).
3a. ZIP model–testing the effect of intervention on the mean of the Poisson part ($H_0 : \beta_1^{ZIP} = 0$).
3b. ZIP model–testing the effect of intervention on the structural zero part ($H_0 : \gamma_1^{ZIP} = 0$).
4. MZIP model–testing the effect of intervention on the overall mean of the entire population ($H_0 : \beta_1^{MZIP} = 0$).
5a. Linear model with raw scale scores–testing the effect of intervention on the overall mean of the entire population ($H_0 : \beta_1^{linear\_raw}$, where $\beta_1^{linear\_raw}$ is the intervention effect in the model).
5b. Linear model with log-transformed outcome scores–testing the effect of intervention on the overall mean of the entire population ($H_0 : \beta_1^{linear\_log}$, where $\beta_1^{linear\_log}$ is the intervention effect in the model). Note that a constant of 1 was added to outcome values to avoid taking the logarithm of zero.

Note that the ZIP model evaluates the intervention effect in two distinct parts, which come with two separate statistical tests for the intervention effect (ie, Models 3a & 3b).

3.5 | Performance measure

To evaluate and compare the performance of considered methods for testing the null hypothesis of H_0: *the intervention has no effect on alcohol consumption*, we calculated the empirical Type I error rate and statistical power of each method. In each simulation setting, we considered 8,000 replications. In each replication, study data were generated based on the data-generating mechanisms explained above, and the considered models were fit to the generated data. After 8,000 replications per condition, we calculated the rejection rate, which is the proportion of replications that yielded statistically significant results for the intervention for each condition for each model. In Conditions 1–4, which have at least one true intervention effect, the observed rejection rate is the empirical statistical power (ie, true positives). In the final Condition 5 with null intervention effects, the rejection rate is the empirical Type I error rate, which is the ability to control the probability of having false positive results. We set the significance level at 0.05, so the rejection rate of 0.05 means that the method adequately controlled the Type I error. If the rejection rate is less than 0.05, the method is overly conservative in controlling the Type I error, which could lead to the inability to detect a true intervention effect, when it exists. If the rejection rate is greater than 0.05, the method may be prone to false positive results. The relative performance across the methods was evaluated by comparing their rejection rates from the models in each of the simulation settings.

4 | RESULTS

4.1 | Type I error

Figure 1 presents the rejection rates of the five statistical methods under null effects (Condition 5) across simulation settings. Since the intervention has no effect on the count and zero parts ($\beta_1 = \gamma_1 = 0$), the rejection rates represent the empirical Type I error. To account for the Monte-Carlo error incurred due to using a finite number of replicates (ie, 8,000), we considered the estimated Type I error to be significantly different than 0.05 if it was outside the range of $\left[0.05 \pm 1.96 \times \sqrt{\frac{0.05*(1-0.05)}{8000}}\right] = [0.045, 0.055]$. From the results, we see that first, the Poisson model failed to control the Type I error rate, which was highly inflated (> 0.06) across simulation settings. Notably, as the zero rate increased, the inflation of the Type I error rate became more severe, indicating that the Poisson model is increasingly likely to lead to false conclusions. Second, the MZIP model, the ZIP model testing the count part (noted as “3a. ZIP: count” in figure legends), and both of the linear models controlled the Type I error rate well (ie, within the expected range $[0.045, 0.055]$) across simulation settings. Third, the ZIP model testing the intervention effect on the structural zero part (noted as “3b. ZIP: zero” in figure legends) controlled Type I error well when sufficient zero inflation existed (> 0.4 zero rate) and the sample size was relatively large (≥ 300). However, when the zero inflation was modest (ie, ≤ 0.4 zero rate) and the sample size was small to modest (< 300), the tests became overly conservative with the Type I error rate less than 0.045. Fourth, the NB model appropriately controlled Type I error when the zero rate was low to moderate (≤ 0.3). However, when the zero rate increased to 0.4 or higher, the Type I error rate generally fell below 0.045. Although the NB model is still valid in terms of statistical significance and its ability to control the probability of false positives, it is more likely to result in excessive false negative results under zero inflation.

4.2 | Statistical power

The rejection rates for the statistical methods under Conditions 1–4 are presented in Figures 2–5, respectively. Since the intervention had effects on at least the count or zero parts (β_1 or $\gamma_1 \neq 0$), the rejection rates represent empirical statistical power. The comparative results between the statistical methods for each of Conditions 1–4 are described. Note that the Poisson model will not be discussed here because it is statistically invalid under zero inflation (Figure 1).

4.2.1 | Condition 1: $\beta_1 \in \{-0.1, -0.2, -0.3\}$ and $\gamma_1 = 0.5$

In Condition 1, the intervention is efficacious for both the count and zero parts. First, as shown in Figure 2, the MZIP model testing the intervention effect on the overall mean and the linear model with raw scale scores had comparable statistical power, which was generally the highest. Second, the linear model on log transformed scores was less powerful

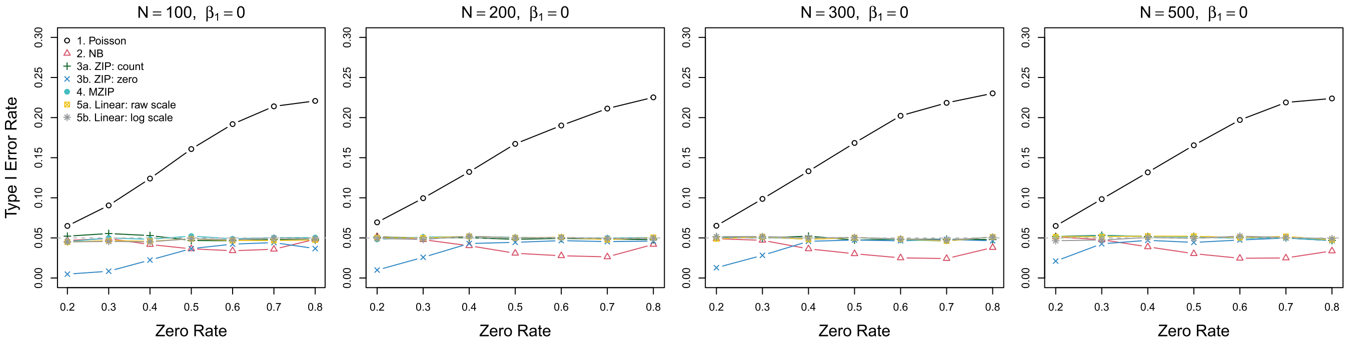

FIGURE 1 Results of empirical Type I error rates under Condition 5 for different statistical methods from 8,000 replications.

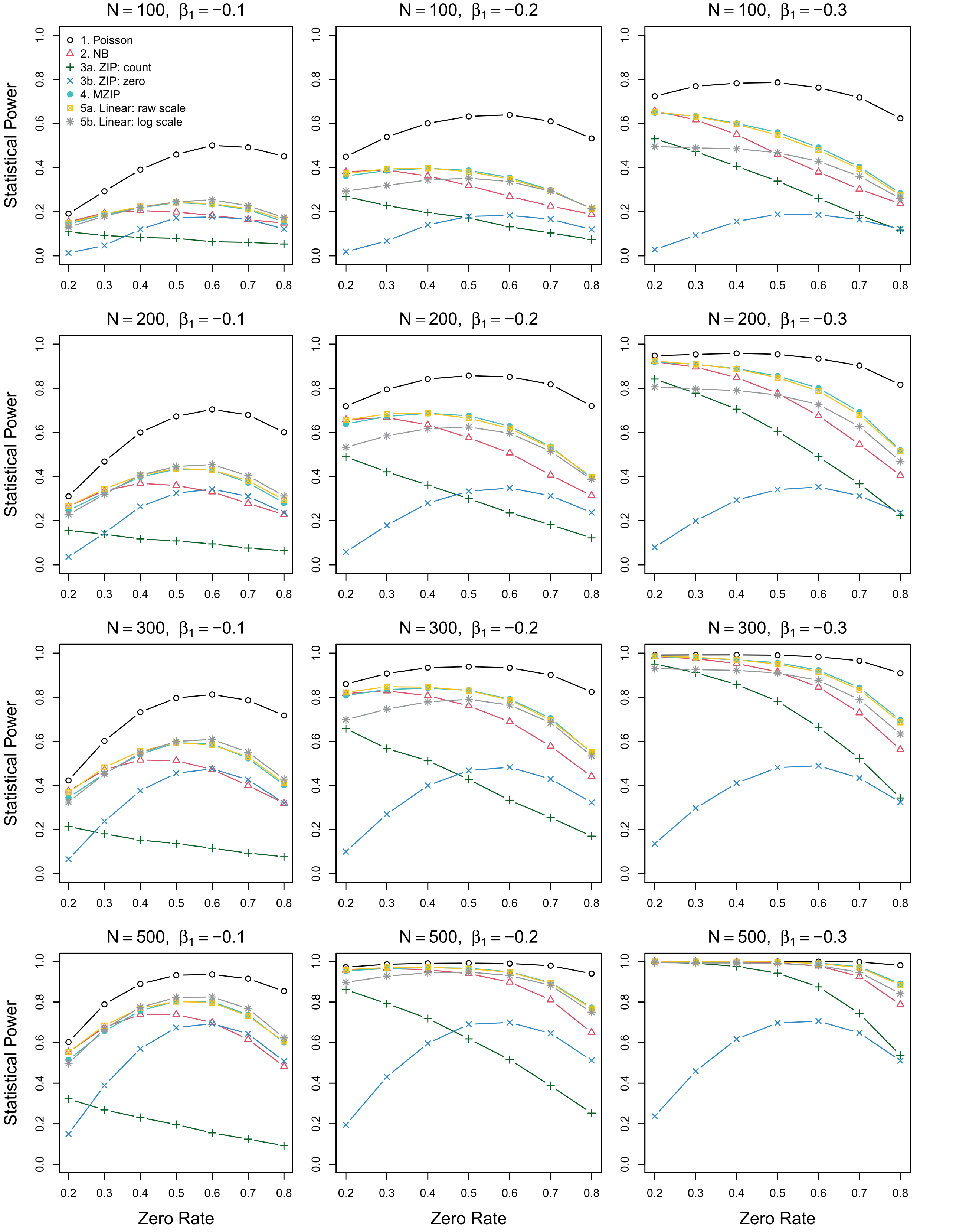

FIGURE 2 Results of empirical statistical power under Condition 1 for different statistical methods from 8,000 replications.

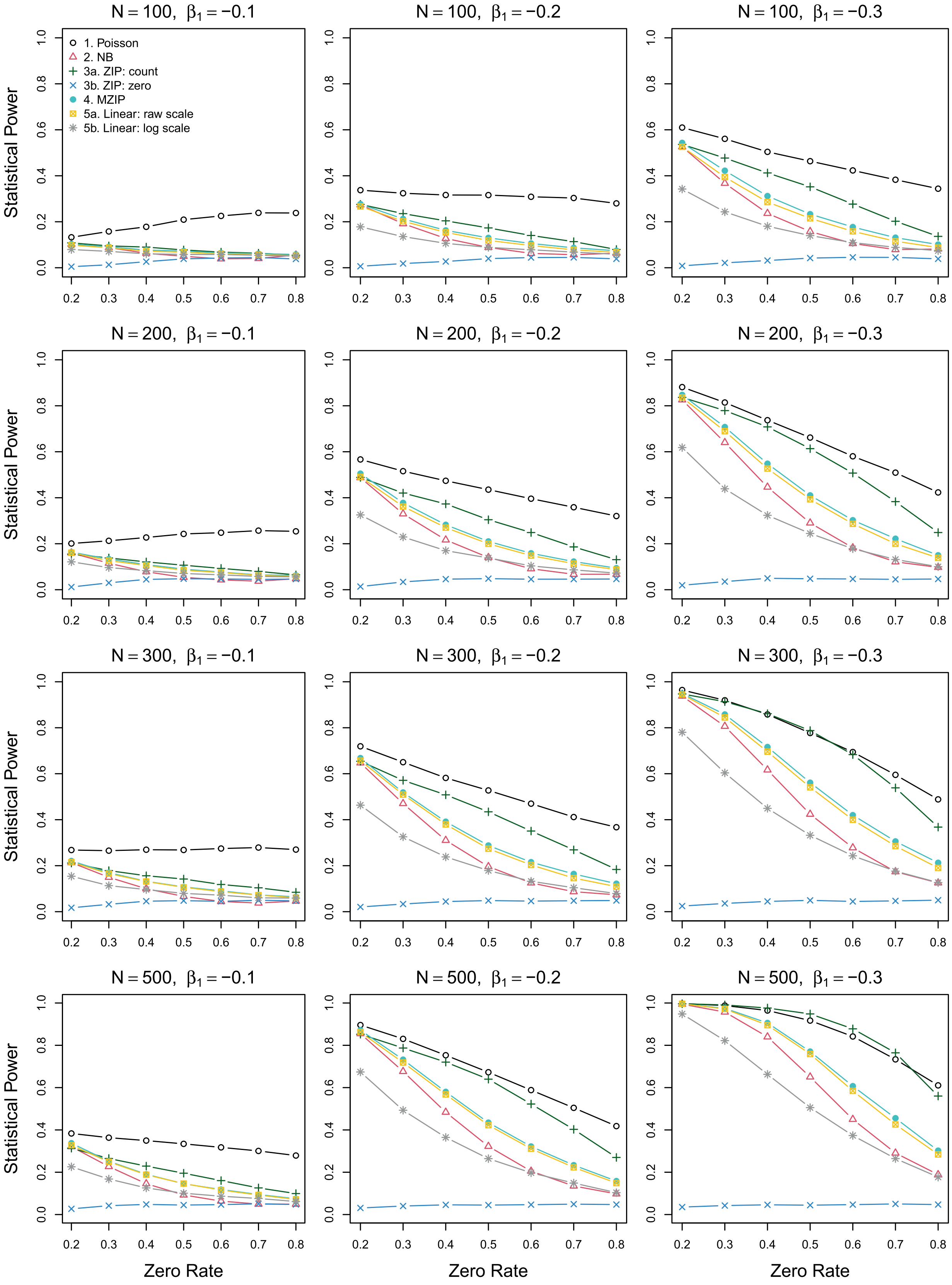

FIGURE 3 Results of empirical statistical power under Condition 2 for different statistical methods from 8,000 replications.

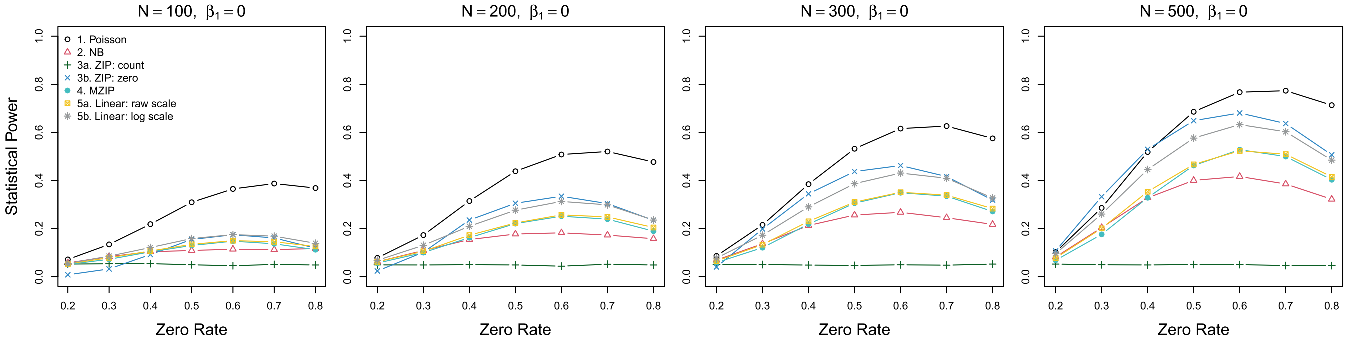

FIGURE 4 Results of empirical statistical power under Condition 3 for different statistical methods from 8,000 replications.

than the MZIP and the linear model with raw scores under higher effect size ($\beta_1 = -0.2$ and -0.3), but the power disadvantage diminished at greater zero rates. When the effect size was small ($\beta_1 = -0.1$), the linear model on log transformed scores had the highest power at high zero rates (> 0.5), but the power gain was modest compared to the MZIP model testing the overall mean and the linear model on the raw scale. Third, the NB model performed well at lower zero rates (≤ 0.3). However, as the zero rates increased, the statistical power of the NB model deteriorated quickly, showing much lower rejection rates than the MZIP or linear models with either raw or log-transformed scores. This observation was expected because of the overly conservative Type I error rate of the NB model under moderate to high levels of zero inflation, which compromised its power to detect true intervention effects. Fourth, the ZIP model testing the count part had less power since testing the intervention effect only for the count part of the population leaves the intervention effects on the zero mass left ignored, leading to power loss.

Notably, under a sample size of $N = 100$, none of the methods reached a power of 0.8 in all simulation conditions. When $N = 200$, a power of 0.8 was only achieved by the MZIP model, linear models, and NB models when the intervention effect on the mean was -0.3 (ie, RR = 0.74) with low to moderate zero inflation. With a sample size of $N = 300$ or greater, statistical power was adequate for the MZIP and linear models in more simulation conditions with a zero rate as a major determining factor of power, along with the magnitude of effects.

4.2.2 | Condition 2: $\beta_1 \in \{-0.1, -0.2, -0.3\}$ and $\gamma_1 = 0$

Suppose that intervention was efficacious only for participants who engaged in the behavior of interest, such as consuming drinks containing alcohol, but not for those who predictably did not drink. First, as shown in Figure 3, the ZIP model testing the Poisson mean outperformed the four other statistical models (ie, MZIP, NB, and linear models with raw or log transformed scores). It may be because the other models could not leverage the intervention effect on the zero mass when estimating the overall treatment effect on the overall mean when $\gamma_1 = 0$, and also because it was the "true" model. Second, the MZIP model testing the intervention effect on the overall mean had the highest power of the remaining four tests, followed by the linear model with raw scores, the NB model, and the linear model with log-transformed scores.

4.2.3 | Condition 3: $\beta_1 = 0$ and $\gamma_1 = 0.5$

In Condition 3, intervention had an effect only on the proportion of zero responses (eg, abstainers). Figure 4 shows the results on the power to detect intervention effects. First, as the "true" model in this condition, the ZIP model testing the structural zero part had the highest power with larger samples ($N \geq 300$). The linear model with log-transformed outcome scores generally had the second highest power, followed by the linear model with raw scores and the MZIP model. Moreover, even at $N = 500$, the power to detect the intervention effect was less than 0.8. For $N = 300$ or less, statistical power was below 0.5. Of note, we anticipate that this condition is less common in practice as interventions generally tend to influence the average number of drinks if they affect the likelihood of abstaining. Therefore, the result of Condition 3 is not our primary interest in the current study.

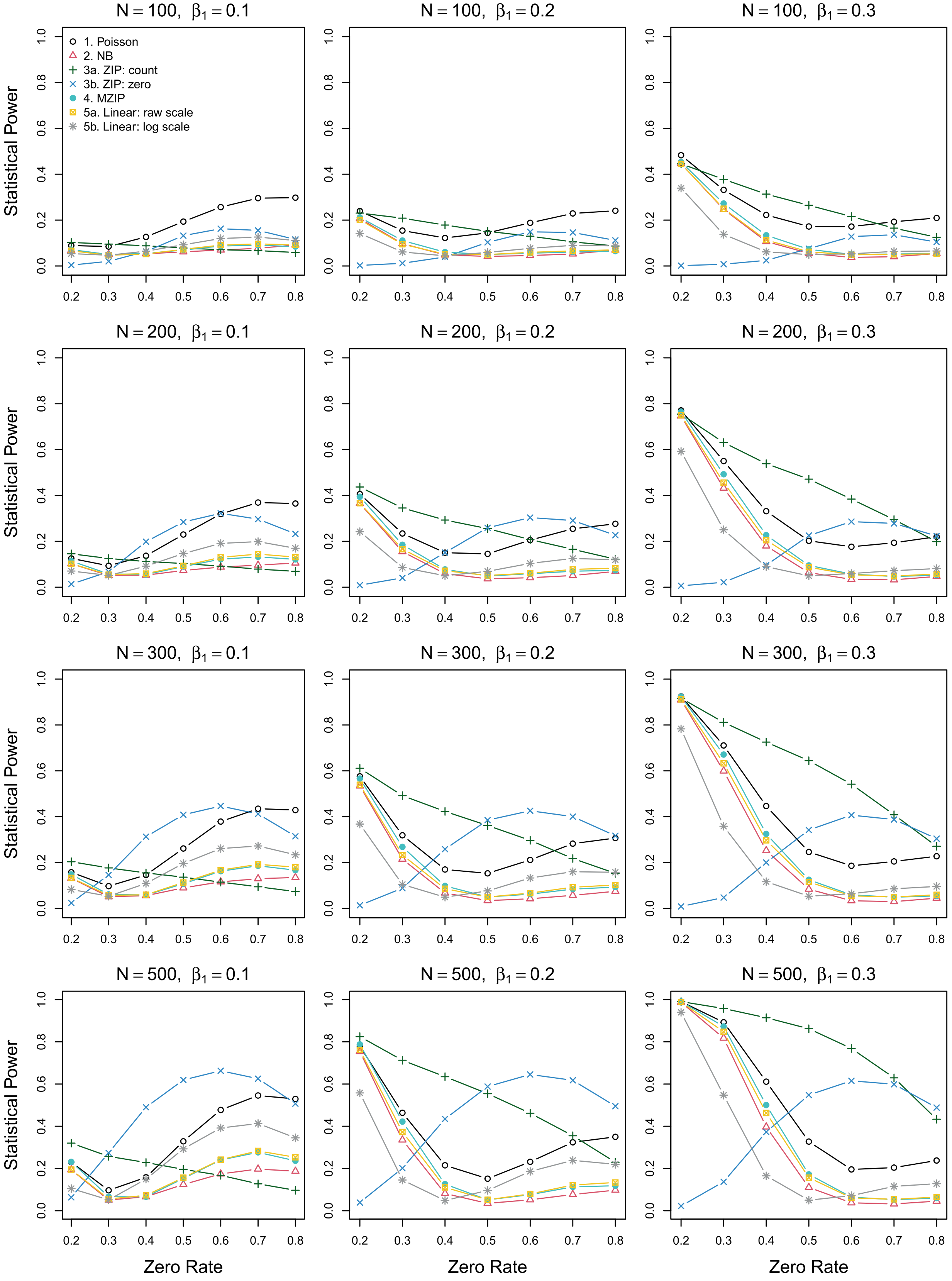

FIGURE 5 Results of empirical statistical power under Condition 4 for different statistical methods from 8,000 replications.

4.2.4 | Condition 4: $\beta_1 \in \{0.1, 0.2, 0.3\}$ and $\gamma_1 = 0.5$

Finally, Condition 4 represents a supposedly less common scenario where the intervention had conflicting effects on the two subpopulations. The empirical power to detect intervention effects is presented in Figure 5. First, the ZIP model testing the count part (3a in figure legends) or zero part (3b in figure legends) generally had the highest power among the methods. Specifically, the test for the count part had higher power as β_1 increased, which was due to a higher effect size, and the test for the zero part reached its highest power when the zero rate was 0.6. Second, the MZIP model generally had a lower power compared to the ZIP model testing the count part and/or the ZIP model testing the zero part, followed by the linear model with raw scores, the NB model, and the linear model with log-transformed scores. Of note, the above four methods focus on the overall mean outcome difference between intervention and control groups. Because the intervention had conflicting effects on the count and zero parts, the intervention effect on the overall mean from one part can be canceled out by the effect on the other part. Clinically speaking, the overall mean number of drinks among the entire sample can be reduced because of the increased number of nondrinkers, which is meaningful. However, any desirable effects on the structural zero part (ie, increased proportion of nondrinkers) may be canceled out or dampened if the average number of drinks among those who drink increases. Therefore, it may be helpful to check the direction of intervention effects on the two parts when deciding which statistical model to select.

4.3 | Computational cost

The computational cost to estimate the methods considered in this study was generally negligible. For example, it took less than 1 second to estimate each considered method for a setting of $N = 500$, $\beta_1 = -0.2$, $\gamma_1 = 0.5$, and zero rate = 0.5. The above experiment was conducted in a laptop with Intel i7-13700HX CPU (2.10 GHz).

5 | DISCUSSION AND CONCLUSIONS

We conducted a methodological phase III study[12] to evaluate and compare the statistical properties of candidate methods for count data through extensive simulation experiments in the context of health behaviors research. This was a neutral comparison study aimed at evaluating and summarizing the relative performance across the considered methods in practical application settings for biostatisticians and applied researchers. We represented a variety of plausible data situations in terms of the size and direction of the intervention effect, sample size, and degree of zero inflation. The empirical results obtained from this simulation can serve to guide real data applications in the field. We provide clinical implications and practical recommendations for model selection.

Among the conventional count distribution-based models without adjustment for zero-inflation, the Poisson model is always invalid with an inflated Type I error rate under zero inflation according to our observation (eg, zero rate ≥ 0.2). The Poisson model tends to falsely judge an ineffective intervention to be efficacious, leading to excessive false positive results. This result can be expected because the Poisson model does not allow modeling excessive zeros nor overdispersion, which typically occurs in zero-inflated data.[22] Therefore, we recommend against using the Poisson model in all scenarios where zero inflation is present. Similar to the Poisson model, the NB model does not account for excessive zeroes, but allows for overdispersion. However, it controls the Type I error below the nominal level in the simulation (eg, $\alpha = 0.05$). When zero inflation is moderate to high (eg, zero rates ≥ 0.4), the NB model tends to overly control Type I error, hampering one's ability to detect true intervention effects. Although the NB model is still statistically valid under zero inflation, its statistical power is compromised.

The ZIP model is an extension of the Poisson model that accounts for excessive zeros through a mixture distribution of the Poisson and a point mass at zero and is most powerful when the main interest is any single part of the ZIP distribution. However, when the overall intervention effect on the entire population is of interest, it has less power than the MZIP model according to our observation. When the intervention has favorable effects on both the subpopulations (ie, reducing the average number of drinks among those who may drink and increasing the proportion of abstainers), the MZIP model generally had superior statistical power in the simulation. This is because the MZIP model evaluates the effects on the overall mean of the outcome directly, which combines the effects from both subpopulations, yielding higher statistical power. Of note, although the ZIP model was used to simulate the data, it may have less power because the tested parameters, β_1^{ZIP} and γ_1^{ZIP}, captured the information from one of the corresponding subpopulations. In contrast, the MZIP and other models were able to capture the information from the entire population. When the intervention has conflicting effects on each of the subpopulations (eg, increasing the average number of drinks among those who drink and also increasing the proportion of abstainers, as shown in Condition 4), the MZIP model underperformed. This is because the intervention effect on the overall mean is attenuated by the effect on one subpopulation canceling out the effect of the other, when the two effects are in opposite directions. However, this illustrates a scenario that is less likely to occur in practice as interventions and treatments are supposed to do no harm to participants.

Under the studied simulation conditions, linear models are also valid with well-controlled Type I error rates. The linear model with raw scale scores generally had higher statistical power compared with its counterpart with log transformation. This observation suggests that although the use of count distribution-based models has been widely promoted for count data, the linear models may still produce valid inferences with acceptable statistical power, especially when compared with the Poisson or NB models. However, the use of log transformation may not be optimal for count outcome analysis with zero inflation due to information loss. For example, the count outcome of number of drinks consumed in a week has a relatively limited range (eg, 0–30 drinks), so taking the log transformation may not be as beneficial as in other situations with a large range, such as expenses in dollars in economic research. Of note, the linear model with raw scale scores had almost identical statistical power to the MZIP model, which may be because both methods target the mean difference in the outcomes. Other considerations in model selection between these two models include the implementation and interpretation of the methods. The linear model is simpler to conduct and is available in nearly all statistical software. The MZIP model is less available in current software, but we have developed an R package "mcount" that facilitates the implementation of this method. Moreover, the linear model assumes the sample to be drawn from a homogeneous population and evaluates the effect based on a linear relationship, while the MZIP assumes two separate subpopulations and evaluates the effect based on a log scale for the overall mean.

This simulation study generated empirical evidence regarding the statistical power and Type I error across candidate methods, which can be useful when selecting an appropriate method in practice. Beyond the relative statistical performance, we recommend applied researchers consider other aspects, such as the underlying data generating mechanism and the appropriate research questions for their study. For example, if it is assumed that a sample was drawn from a homogeneous population, then statistical methods based on a mixture distribution (eg, ZIP and MZIP models) may not be advisable. Another example is, if it is reasonable to assume there exists two subpopulations (eg, those at-risk and not-at-risk) and questions of interest concern only on one of the subpopulations (eg, intervention aims to reduce alcohol use among those who may drink), then the ZIP model may be preferred.

The simulation settings studied in the current study were motivated by the data from brief alcohol intervention studies. Therefore, the findings are most relevant for the alcohol intervention trials with similar effect sizes. With zero-inflated outcome data, we also observed that statistical power across the methods was mostly below 0.8 under small to moderate sample sizes (eg, $N \leq 300$) and effect sizes in the simulation (eg, $|\beta_1| \leq 0.2$, corresponding to an intervention effect that reduces no more than 19% of the average number of drinks among participants who may drink). Even at a large sample size of $N = 500$, the power was adequate (at the 0.8 level) only for a few conditions in the simulation. Specifically, typical clinical trials would not have adequate power to detect a significant intervention effect under small to moderate effect sizes with small to moderate samples, regardless of the statistical methods used. Our findings indicate that in the presence of zero inflation, individual clinical trials lack statistical power to detect effects on count outcomes. This underscores the value of meta-analysis using individual participant data for increasing statistical power when analyzing such data (eg, Reference 9). Study-specific small or null effects can be combined together to provide large-scale robust evidence in the field of brief alcohol intervention and related areas (eg, Reference 21).

AUTHOR CONTRIBUTIONS

ZZ, DL, and EYM initially wrote the manuscript's early drafts. ZZ and DL prepared the computing script, and ZZ conducted the simulation and prepared all figures and tables. DH helped conceptualize a methodological gap in clinical applications. EYM and MX provided feedback on the design and interpretation of the simulation study. All authors reviewed and edited earlier drafts and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The project described was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) grants R01 AA019511 and K02 AA028630, the National Institute of Allergy and Infectious Diseases (NIAID) grant R21 AI179404, and the National Science Foundation (NSF) grants DMS2015373, DMS2027855, DMS2311064, and DMS2319260. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA, the NIAID, the National Institutes of Health, or the NSF.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

The program code of the MZIP method is available from the R package "mcount".[13] The script for simulation and analysis is available at Mendeley Data (https://data.mendeley.com/datasets/r5bztdd766/2; Reference 23).

ORCID

Zhengyang Zhou https://orcid.org/0000-0002-8039-418X
Dateng Li https://orcid.org/0000-0002-4287-5337

REFERENCES

1. Huh D, Mun E-Y, Walters ST, Zhou Z, Atkins DC. A tutorial on individual participant data meta-analysis using bayesian multilevel modeling to estimate alcohol intervention effects across heterogeneous studies. *Addict Behav*. 2019;94:162-170.
2. Sheu M-L, Hu T-W, Keeler TE, Ong M, Sung H-Y. The effect of a major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. *Health Econ*. 2004;13(8):781-791.
3. Hutchinson MK, Jemmott JB III, Jemmott LS, Braverman P, Fong GT. The role of mother–daughter sexual risk communication in reducing sexual risk behaviors among urban adolescent females: a prospective study. *J Adolesc Health*. 2003;33(2):98-107.
4. Huh D, Mun E-Y, Larimer ME, et al. Brief motivational interventions for college student drinking may not be as powerful as we think: an individual participant-level data meta-analysis. *Alcohol Clin Exp Res*. 2015;39(5):919-931.
5. Zhou Z, Xie M, Huh D, Mun E-Y. A bias correction method in meta-analysis of randomized clinical trials with no adjustments for zero-inflated outcomes. *Stat Med*. 2021;40(26):5894-5909.
6. Famoye F, Preisser JS. Marginalized zero-inflated generalized Poisson regression. *J Appl Stat*. 2018;45(7):1247-1259.
7. Long DL, Preisser JS, Herring AH, Golin CE. A marginalized zero-inflated Poisson regression model with overall exposure effects. *Stat Med*. 2014;33(29):5151-5165.
8. Preisser JS, Long DL, Stamm JW. Matching the statistical model to the research question for dental caries indices with many zero counts. *Caries Res*. 2017;51(3):198-208.
9. Mun E-Y, Zhou Z, Huh D, et al. Brief alcohol interventions are effective through 6 months: findings from marginalized zero-inflated poisson and negative binomial models in a two-step ipd meta-analysis. *Prev Sci*. 2023;24(8):1608-1621.
10. Zhou Z, Li D, Zhang S. Sample size calculation for cluster randomized trials with zero-inflated count outcomes. *Stat Med*. 2022;41(12):2191-2204.
11. Tan L, Luningham JM, Huh D, et al. The selection of statistical models for reporting count outcomes and intervention effects in brief alcohol intervention trials: a review and recommendations. *Alcohol: Clin Experi Res*. 2024;48(1):16-28.
12. Heinze G, Boulesteix A-L, Kammer M, Morris TP, White IR, STRATOS initiative, S. P. Phases of methodological research in biostatistics—building the evidence base for new methods. *Biom J*. 2024;66(1):2200222.
13. Zhou Z, Li D, Huh D, Mun E-Y. mcount: Marginalized count regression models. 2022. https://CRAN.R-project.org/package=mcount. R package version 1.0.0
14. Boulesteix A-L, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. *PLoS One*. 2013;8(4):e61562.
15. Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med Res Methodol*. 2017;17:138.
16. Hilbe JM. *Negative Binomial Regression*. Cambridge, UK: Cambridge University Press; 2011.
17. Perumean-Chaney SE, Morgan C, McDowall D, Aban I. Zero-inflated and overdispersed: what's one to do? *J Stat Comput Simul*. 2013;83(9):1671-1683.
18. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Dent Tech*. 1992;34(1):1-14.
19. Mullahy J. Specification and testing of some modified count data models. *J Econometr*. 1986;33(3):341-365.
20. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
21. Mun E-Y, De La Torre J, Atkins DC, et al. Project INTEGRATE: an integrative study of brief alcohol interventions for college students. *Psychol Addict Behav*. 2015;29(1):34-48.

22. Yang Z, Hardin JW, Addy CL. Testing overdispersion in the zero-inflated Poisson model. *J Stat Plan Inference*. 2009;139(9): 3340-3353.

23. Zhou Z, Li D, Huh D, Xie M, Mun E-Y. Script for: Zhou et al. (2024). A simulation study of the performance of statistical models for count outcomes with excessive zeros. Mendeley Data, V2. 2024. doi:10.17632/r5bztdd766.2

How to cite this article: Zhou Z, Li D, Huh D, Xie M, Mun E-Y. A simulation study of the performance of statistical models for count outcomes with excessive zeros. *Statistics in Medicine*. 2024;1-16. doi: 10.1002/sim.10198