Received: 15 July 2020 | Accepted: 5 January 2021

DOI: 10.1111/2041-210X.13559

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

The consequences of checking for zero-inflation and overdispersion in the analysis of count data

Harlan Campbell

The University of British Columbia, Vancouver, BC, Canada

Correspondence
Harlan Campbell
Email: harlan.campbell@stat.ubc.ca

Handling Editor: Robert B. O'Hara

Abstract

1. Count data are ubiquitous in ecology and the Poisson generalized linear model (GLM) is commonly used to model the association between counts and explanatory variables of interest. When fitting this model to the data, one typically proceeds by first confirming that the model assumptions are satisfied. If the residuals appear to be overdispersed or if there is zero-inflation, key assumptions of the Poison GLM may be violated and researchers will then typically consider alternatives to the Poison GLM. An important question is whether the potential model selection bias introduced by this data-driven multi-stage procedure merits concern.
2. Here we conduct a large-scale simulation study to investigate the potential consequences of model selection bias that can arise in the simple scenario of analysing a sample of potentially overdispersed, potentially zero-inflated, count data. Specifically, we investigate model selection procedures recently recommended by Blasco-Moreno et al. (2019) using either a series of score tests or information theoretic criteria to select the best model.
3. We find that, when sample sizes are small, model selection based on preliminary score tests (or information theoretic criteria, e.g. AIC, BIC) can lead to potentially substantial inflation of false positive rates (i.e. type 1 error inflation). When sample sizes are sufficiently large, model selection based on preliminary score tests, is not problematic.
4. Ignoring the possibility of overdispersion and zero-inflation during data analyses can lead to invalid inference. However, if one does not have sufficient power to test for overdispersion and zero-inflation, post hoc model selection may also lead to substantial bias. This 'catch-22' suggests that, if sample sizes are small, a healthy skepticism is warranted whenever one rejects the null hypothesis of no association between a given outcome and covariate.

KEYWORDS

model selection bias, overdispersion, zero-inflated models, zero-inflation

1 | INTRODUCTION

Despite the ongoing debate surrounding the use (and misuse) of significance testing in ecology (Dushoff et al., 2019; Murtaugh, 2014) and in other fields (Amrhein et al., 2019), hypothesis testing remains prevalent. Indeed, many research fields have been criticized for publishing studies with serious errors of testing and interpretation, and ecologists have been accused of being 'confused' about

when and how to conduct appropriate hypothesis tests (Stephens et al., 2005). One issue that receives a substantial amount of attention is that of failing to check for possible violations of distributional assumptions. According to Freckleton (2009), using statistical tests that assume a given distribution on the data while failing to test for the assumptions required of said distribution is one of 'seven deadly sins'.

One of the most popular statistical models in ecology (and in many other fields, e.g. finance, psychology, neuroscience, and microbiome research (Bening & Korolev, 2012, Loeys et al., 2012, Zoltowski & Pillow, 2018, Xu et al., 2015)) is the Poisson generalized linear model (GLM; Nelder & Wedderburn, 1972). With count response data, a Poisson GLM is the most common starting point for testing an association between a given response, *Y*, and a given covariate of interest, *X*. A Poisson GLM assumes the response data, conditional on the covariates, are the result of independent sampling from a Poisson distribution where, importantly, the mean and variance are equal. However, in practice, count data will often show more variation than is implied by the Poisson distribution and the use of Poisson models is not always appropriate (Cox, 1983).

Count data frequently exhibit two (related) characteristics: (a) overdispersion and (b) zero-inflation. Overdispersion refers to an excess of variability, while zero-inflation refers to an excess of zeros (Yang et al., 2010). If model residuals are overdispersed or have an excess of zeros, assumptions underlying a Poisson GLM will not hold and ignoring this will lead to serious errors (e.g. biased parameter estimates and invalid standard errors; Harrison, 2014). It is therefore routine practice for researchers to check if the assumptions required of a Poisson model hold and adopt an alternative statistical model in the event that they do not; see Zuur et al. (2010).

In the case of overdispersion, popular alternatives to the Poison GLM include the Quasi-Poisson (QP) model (Wedderburn, 1974) and the negative binomial (NB) model (Lindén & Mäntyniemi, 2011; Richards, 2008). (Note that when selecting between the QP and NB models, the best choice is not always straightforward; see Ver Hoef and Boveng (2007), also see Potts and Elith (2006).) In the case of zero-inflation, popular alternatives to the Poison GLM include the zero-inflated Poisson model (ZIP; Lambert, 1992; Martin et al., 2005) and the zero-inflated negative binomial model (ZINB; Greene, 1994).

A multi-stage procedure will typically have researchers testing for overdispersion and zero-inflation in a preliminary stage (based on the residuals from a Poisson GLM), before testing the main hypothesis of interest (i.e. the association between *Y* and *X*) in a second stage; see Blasco-Moreno et al. (2019). If the first stage tests are not significant, a Poisson GLM is selected, regression coefficients are estimated along with their standard errors, and *p*-values are calculated allowing one to test for the association between *Y* and *X*. On the other hand, if the first stage test for overdispersion is significant, a QP or a NB model will be fit to the data. Or, alternatively, if the first stage test for zero-inflation is significant, a ZIP model may be used. In cases when there is evidence of both overdispersion and zero-inflation, more complex models such as the ZINB model or hurdle models will often be considered; see Zorn (1998).

Such a multi-stage, multi-test procedure may appear rather reasonable, and goodness-of-fit tests are frequently reported to confirm that the model-selection is appropriate. However, recently, some researchers have warned against preliminary testing for distributional assumptions; for example, Shuster (2005) and Wells and Hintze (2007). Their warnings are based on the following concern: since the preliminary tests are applied to the same data as the main hypothesis tests, this multi-stage procedure amounts to 'using the data twice' or 'double dipping'; see Devezer et al. (2020) and Kriegeskorte et al. (2009). A hypothesis test using a model selected based on preliminary testing fails to take into account one's uncertainty with regards to the distributional properties of the data. Unless the preliminary tests and the main hypothesis tests are entirely independent, this can result in model selection bias.

The model selection bias at issue here is not the better known model selection bias associated with deciding post hoc which variables to include in the model, for example, the model selection bias associated with stepwise regression (Hurvich & Tsai, 1990; Whittingham et al., 2005). Instead, here we are concerned with the potential bias introduced when deciding post hoc which distributional assumptions should be accepted. The implications of considering post hoc alternatives (or adjustments) to accommodate for distributional assumptions have been previously considered in other contexts. Three examples come to mind.

First, in the context of time-to-event data, the consequences of checking and adjusting for potential violations of the proportional hazards (PH) assumption required of a Cox PH model are considered by Campbell and Dean (2014). The authors find that the 'common two-stage approach' (in which one selects a model based on a preliminary test for PH) can lead to substantial inflation of false-positive rates (i.e. inflation of the type 1 error), even in scenarios where there is no violation of the PH assumption.

Second, in the simple context of testing the means of two independent samples, Rochon et al. (2012) investigate the consequences of conducting a preliminary test for normality (e.g. the Shapiro–Wilk test). The authors conclude that while '[f]rom a formal perspective, preliminary testing for normality is incorrect and should therefore be avoided', in practice, 'preliminary testing does not seem to cause much harm'.

Finally, in the context of clinical trials, factorial trials are an efficient method of estimating multiple treatments in a single trial. However, factorial trials rely on the strict assumption of no interaction between the different treatments. Kahan (2013) investigates the consequences of conducting a preliminary test for the interaction between treatment arms (as is often recommended). By means of a simulation study, Kahan (2013) shows that the estimated treatment effect from a factorial trial under the 'two-stage analysis' can be severely biased, even in the absence of a true interaction.

Model averaging is a possible solution to the problem of model-selection bias. However, model averaging is known to be computationally demanding and correctly interpreting parameter estimates may be difficult; see Hooten and Hefley (2019). Indeed, Cade (2015) warns of 'seriously compromised statistical interpretations'.

Model selection bias is considered a 'quiet scandal in the statistical community' (Breiman, 1992) and is now all the more important to understand given recent concerns with research reproducibility and researcher incentives (Campbell & Gustafson, 2019; Fraser et al., 2018; Gelman & Loken, 2013; Kelly, 2019; Nosek et al., 2012). In some fields, such as psychology, the issue is finally being recognized. Williams and Albers (2019) conclude that 'it is currently unclear how [psychology] researchers should deal with distributional assumptions' since 'diagnosing and responding to distributional assumption problems' may result in 'error rates [that] vary considerably from the nominal error rates'.

In ecology, some have warned about model selection bias (e.g. Buckland et al., 1997), but the problem 'remains widely over-looked' (Whittingham et al., 2006). Indeed, ecologists will readily admit that 'this problem is commonly not appreciated in modelling applications' (Whittingham et al., 2005). Anderson (2007) notes that: 'Model selection bias is subtle but its effects are widespread and little understood by many people working in the life sciences'.

In this paper, we conduct a large-scale simulation study to investigate the potential consequences of model selection bias that can arise in the simple scenario of analysing a sample of potentially overdispersed, potentially zero-inflated, count data. It is difficult to anticipate what these consequences might be. Often, while model selection bias is problematic from a theoretical perspective, it does not lead to substantial problems in practice. We restrict our attention to two model selection procedures, one based on conducting score tests, and another based on calculating information criteria. These correspond to recommendations recently put forth in Blasco-Moreno et al. (2019).

In Section 2, we review commonly used models and outline the framework of a simulation study to investigate the consequences of checking for zero-inflation and overdispersion. In Section 3, we discuss the results of this simulation study and we conclude in Section 4 with a summary of findings and general recommendations.

2 | MATERIALS AND METHODS

2.1 | Models for the analysis of count data

Let us begin with the simplest version of the Poisson GLM. Let Y_i, for i in 1, ..., n, be the response of interest observed for n independent samples. Let X_i, for i in 1, ..., n, represent a single covariate of interest. If the covariate of interest is categorical with k different categories (e.g. k different species of fish), X_i will be a vector with length equal to $k - 1$; otherwise it will be a single scalar (and $k = 2$). The simplest Poisson regression model, with a standard log link, will have:

$$Y_i \sim \text{Poisson}\left(\lambda_i = \exp\left(\beta_0 + \beta_X X_i\right)\right), \quad \text{or equivalently:} \tag{1}$$

$$\Pr\left(Y_i = y_i | \beta_0, \beta_X\right) = \frac{\left(\exp\left(\beta_0 + \beta_X X_i\right)\right)^{y_i} \exp\left(-\exp\left(\beta_0 + \beta_X X_i\right)\right)}{y_i!}, \tag{2}$$

for i in 1, ..., n, where β_0 is the intercept, and β_X is the coefficient (or coefficient-vector of length $k - 1$) representing the association between X and Y. Note that this model assumes that the mean and variance are equal: $\text{E}(Y_i) = \text{Var}(Y_i) = \lambda_i$, for i in 1, ..., n.

Parameter estimates, $\widehat{\beta_0}$, and $\widehat{\beta_X}$, can be obtained by maximum likelihood estimation. A confidence interval for β_X is typically calculated by the standard profile likelihood approach where one inverts a likelihood-ratio test (LRT); see Venzon and Moolgavkar (1988), or more recently Uusipaikka (2008).

To test whether there is an association between Y and X, we define the following hypothesis test: $H_0 : \beta_X = 0$ versus $H_1 : \beta_X \neq 0$. A simple LRT, or Wald test will provide a p-value to evaluate this hypothesis; see Zeileis et al. (2008). The LRT and Wald test are asymptotically equivalent. For the LRT, the Z-statistic is obtained by calculating the null and residual deviance as $Z_{\text{LRT}} = D_1 - D_0$, where:

$$D_0 = 2\sum_{i=1}^{n}\left\{Y_i \log\left(Y_i / \exp\left(\widehat{\beta}_0\right)\right) - \left(Y_i - \exp\left(\widehat{\beta}_0\right)\right)\right\},$$

and:

$$D_1 = 2\sum_{i=1}^{n}\left\{Y_i \log\left(Y_i / \widehat{\lambda}_i\right) - \left(Y_i - \widehat{\lambda}_i\right)\right\}, \quad \text{where } \widehat{\lambda}_i = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_X X_i\right).$$

If the distributional assumptions of a Poisson GLM are met and the null hypothesis holds, the Z-statistic will follow (asymptotically) a χ^2 distribution with $df = k - 1$ degrees of freedom, and the p-value is calculated as: $p\text{-value} = P_{\chi 2}(Z, df = k - 1)$. (For the Wald test, with $k = 2$, the Z-statistic is defined as $Z_{\text{Wald}} = \left(\widehat{\beta_X} / se\left(\widehat{\beta_X}\right)\right)^2$, where $se\left(\widehat{\beta_X}\right)$ is the standard error of the maximum likelihood estimate (MLE); see Hilbe and Greene (2007) for details when $k > 2$).

However, if the distributional assumptions do not hold, the Z-statistic will be compared with the wrong reference distribution invalidating any significance test (and associated confidence intervals). Therefore, in order to conduct valid inference, researchers will typically carry out an extensive model selection procedure. Note that model selection must always be based on model residuals and not on the distribution of the response variable (which is erroneously done on occasion). To be clear, one should not check the distribution of the response variable independent of the covariates.

Blasco-Moreno et al. (2019) outline and illustrate a proposed model selection protocol based on:

- measuring indices (e.g. the dispersion index (Fisher, 1950); the zero-ination index (Puig & Valero, 2006));
- conducting score tests (e.g. the *D&L* score test for Poisson vs. NB regression (Dean & Lawless, 1989): the νdB score test for Poisson vs. ZIP regression (Van den Broek, 1995): the score test for ZIP vs. ZINB regression (Ridout et al., 2001)); and
- evaluating candidate models with goodness-of-fit tests (e.g. likelihood ratio tests; the Vuong and Clarke tests) and model selection criteria (e.g. AIC and BIC).

In this paper, for simplicity, we will only consider three alternative models in addition to the Poisson model described above: the

(type 2) NB, the ZIP, and the (type 2) ZINB regression models as described in Blasco-Moreno et al. (2019). Let us briefly review these three alternative regression models.

(1) *The ZIP regression model*—We will consider the following zero-inflated Poisson model where the probability of a structural zero, ω_i, is a function of the covariate X_i. Specifically,

$$\begin{aligned}\Pr(Y_i = y_i|\omega_i, \lambda_i) &= \omega_i + (1-\omega_i)\exp(-\lambda_i), \quad \text{if } y=0;\\ &= (1-\omega_i)\exp(-\lambda_i)\lambda_i^{y_i}/y_i!, \quad \text{if } y_i>0;\end{aligned} \tag{3}$$

where we have a log link function for λ_i and a logit link function for ω_i (for i in 1, …, n) such that:

$$\lambda_i = \exp(\beta_0 + \beta_X X_i), \quad \text{and} \tag{4}$$

$$\omega_i = \left(\frac{\exp(\gamma_0 + \gamma_X X_i)}{1+\exp(\gamma_0 + \gamma_X X_i)}\right). \tag{5}$$

The ZIP model has that $0 \le \omega_i \le 1$ and $\lambda_i > 0$, and implies the following about the mean and variance of the data: $E(Y_i) = \lambda_i(1 - \omega_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \mu_i^2\omega_i/(1-\omega_i)$. The dispersion index is therefore equal to $d = \text{Var}(Y_i)/E(Y_i) = 1 + \lambda_i\omega_i$. As $\omega_i \to 0$, we have that Y_i reverts to follow a Poisson distribution with mean λ_i. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = \gamma_X = 0$.

(2) *The (type 2) NB regression model*—We will consider the following NB regression model:

$$\Pr(Y_i = y_i|\nu, \lambda_i) = \frac{\Gamma(y_i+\nu)}{\Gamma(y_i+1)\Gamma(\nu)}\left(\frac{1}{1+\lambda_i/\nu}\right)^{\nu}\left(\frac{\lambda_i/\nu}{1+\lambda_i/\nu}\right)^{y_i}; \tag{6}$$

where we use a log link function for $\lambda_i = \exp(\beta_0 + \beta_X X_i)$, and where $\nu > 0$ is a dispersion parameter that does not depend on covariates. The type 2 NB model assumes the following about the mean and variance of the data: $E(Y_i) = \lambda_i$, and $\text{Var}(Y_i) = \lambda_i + \lambda_i^2/\nu$. The dispersion index is therefore equal to $d = \text{Var}(Y_i)/E(Y_i) = 1 + \lambda_i/\nu$. As $\nu \to \infty$, we have that Y_i reverts to follow a Poisson distribution with mean λ_i. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = 0$.

(3) *The (type 2) ZINB regression model*—We will consider the following ZINB regression model:

$$\begin{aligned}\Pr(Y_i = y_i|\nu, \omega_i, \lambda_i) &= \omega_i + (1-\omega_i)(1/(1+\lambda_i/\nu))^{\nu}, \quad \text{if } y=0;\\ &= (1-\omega_i)\frac{\Gamma(y_i+\nu)}{\Gamma(y_i+1)\Gamma(\nu)}\left(\frac{1}{1+\lambda_i/\nu}\right)^{\nu}\left(\frac{\lambda_i/\nu}{1+\lambda_i/\nu}\right)^{y_i}, \quad \text{if } y_i>0;\end{aligned} \tag{7}$$

where we use a log link function for λ_i and a logit link function for ω_i as described in Equations 4 and 5; and where $\nu > 0$ is a dispersion parameter that does not depend on covariates. The type 2 ZINB model assumes the following about the mean and variance of the data: $E(Y_i) = \lambda_i(1 - \omega_i)$, and $\text{Var}(Y_i) = (1-\omega_i)(\lambda_i + \lambda_i^2(\omega_i + 1/\nu))$. The dispersion index is therefore equal to $d = \text{Var}(Y_i)/E(Y_i) = 1 + \lambda_i(\omega_i + 1/\nu)$. A null hypothesis of no association between X and Y is specified as: $H_0 : \beta_X = \gamma_X = 0$.

2.2 | Simulation study

As discussed in the previous section, prevailing practice for the analysis of count data is first to try to fit one's data with a Poisson GLM and only consider alternatives in the event that a preliminary test indicates that the distributional assumptions of a Poisson GLM may be violated. We will therefore consider the following multi-stage testing procedure in our simulation study investigation. This follows the recommendations of Blasco-Moreno et al. (2019) yet represents a simplication of the typical process followed by researchers. (Walters (2007) also recommends a similar multi-step model selection procedure.)

For the illustrative purposes of this paper, we consider the Dean and Lawless (1989) score test (*D&L* test) for oversdispersion and the Vuong (1989) test for zero-inflation (see Appendix S1 for details) in the following seven step procedure:

Step 1. Conduct the *D&L* score test for overdispersion (H_0: Poisson vs. H_1: NB).

Step 2. If the *D&L* score test fails to reject the null, conduct a Vuong test for zero-inflation (H_0: Poisson vs. H_1: ZIP). Otherwise, proceed to Step 5.

Step 3. If the Vuong test for zero-inflation fails to reject the null, fit a Poisson GLM and calculate the p-value ($H_0 : \beta_X = 0$ vs. $H_1 : \beta_X \neq 0$). Otherwise, proceed to Step 4.

Step 4. If the Vuong test for zero-inflation rejects the null, fit the ZIP model and calculate the p-value ($H_0 : \beta_X = \gamma_X = 0$).

Step 5. If the *D&L* score test rejects the null, conduct the Vuong test for zero-inflation (H_0: NB vs. H_1: ZINB).

Step 6. If the Vuong test for zero-inflation fails to reject the null, fit the NB model and calculate the p-value ($H_0 : \beta_X = 0$). Otherwise, proceed to Step 7.

Step 7. If the Vuong test for zero-inflation rejects the null, fit the ZINB regression model and calculate the p-value ($H_0 : \beta_X = \gamma_X = 0$).

Figure 1 illustrates the multi-stage model selection procedure with a Poisson GLM as the starting point. Note that, in their example analysis of plant–herbivore interaction data, Blasco-Moreno et al. (2019) conduct a version of the above procedure. First, based on the *D&L* score test, they conclude: 'the data are clearly overdispersed and a NB model was preferred to a Poisson'. The authors also conduct Vuong and Clarke tests: 'The Vuong and Clarke tests rejected the Poisson and NB models in favour of their zero-inflated versions[…]'. We decided to consider the Vuong test in our simulations instead of the Clarke test (or the Ridout score test), since the Vuong test appears to be the most widely used in practice. We also investigate two other, simpler, model selection strategies: (a) among the four models considered, the model with lowest AIC is chosen; (b) among the four models considered, the model with lowest BIC is chosen (Burnham & Anderson, 2004). And we also briefly consider a third alternative: among the four models considered, the model with lowest AICc is chosen (Hurvich & Tsai, 1989). Different information

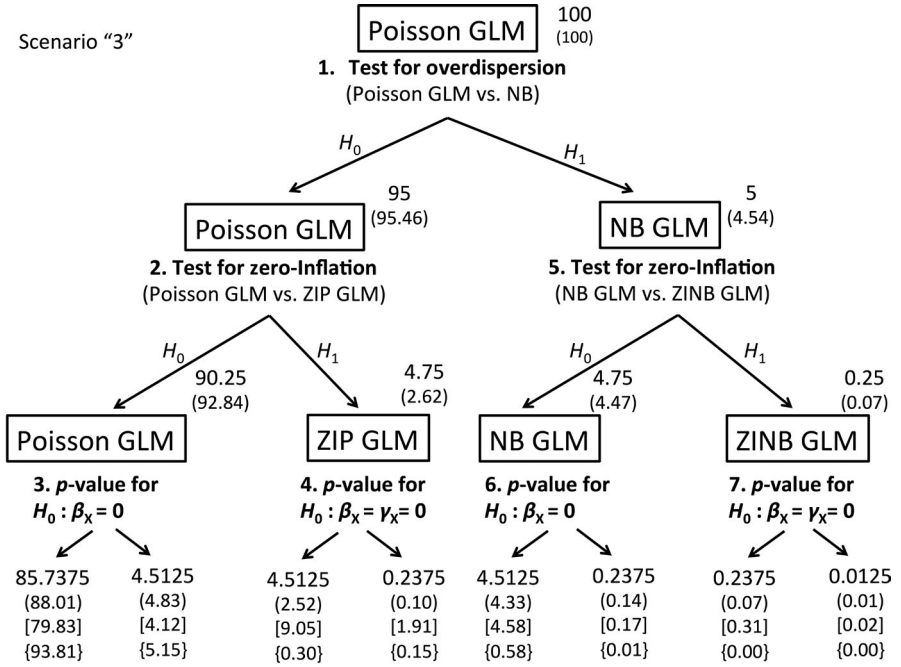

FIGURE 1 The multi-stage model selection procedure. A Poisson GLM is the starting point. Three score tests lead to one of four models. Numbers in the top right-hand corner of each node indicate the expected number of datasets (out of a total of 100) to reach each outcome if the data were Poisson (with $\beta_X = 0$), and each of the tests were truly independent (with a $\alpha = 0.05$ false positive rate). Numbers in parentheses correspond to results from the simulation study (Scenario '3', with $\phi = \infty$, $\omega = 0$, $\beta_0 = 0.5$ and $n = 250$). Numbers in (curved) parentheses are those obtained in following the seven-step procedure; numbers in [square] parentheses are those obtained via AIC; numbers in {curly} parentheses are those obtained via BIC. The unconditional type 1 error rate obtained in the simulation study following the seven-step procedure is 5.08% (=4.83 + 0.10 + 0.14 + 0.01). The unconditional type 1 error rate obtained by the simulation study when selecting the best model via AIC is 6.23% (=4.12 + 1.91 + 0.17 + 0.02). (Not plotted, but for reference: the unconditional type 1 error rate obtained by the simulation study when selecting the best model via AICc is 6.21% (=4.21 + 1.83 + 0.16 + 0.01)) The unconditional type 1 error rate obtained the simulation study when selecting the best model via BIC is 5.31% (=5.15 + 0.15 + 0.01 + 0.00)

criteria are known to have different properties. For instance, AIC is optimal for reducing predictive error, whereas the BIC is consistent; see Yang (2005).

We conducted a large-scale simulation study in which samples of data were drawn from four different distributions (see R code in the Appendix S1 for exactly how the data simulation is done):

1. the Poisson distribution:
   $y_i \sim \text{Poisson}\left(\lambda = \exp\left(\beta_0\right)\right)$, for i in $1, \ldots, n$;
2. the (type 2) negative binomial distribution:
   $y_i \sim \text{NegBin}\left(\nu, \lambda = \exp\left(\beta_0\right)\right)$, for i in $1, \ldots, n$;
3. the zero-inflated Poisson distribution:
   $y_i \sim \text{ZIPoisson}\left(\omega, \lambda = \exp\left(\beta_0\right)\right)$, for i in $1, \ldots, n$; and
4. the zero-inflated negative binomial distribution:
   $y_i \sim \text{ZINegBin}\left(\nu, \omega, \lambda = \exp\left(\beta_0\right)\right)$, for i in $1, \ldots, n$.

For each scenario, all data are simulated under the null hypothesis (i.e. with $\beta_X = 0$ and $\gamma_X = 0$). We varied the following: the sample size, n = (50, 100, 250, 500, 1,000, 2,000, 5,000, 10,000), the intercept, β_0 = (0.5, 1.0, 1.5, 2.0, 2.5), and the probability of a structural zero, ω = (0, 0.05, 0.1, 0.2, 0.5). We also varied the degree of overdispersion by setting $\phi = \nu/\lambda$ = (∞, 2, 1, 1/2, 1/3) (so that data simulated from the negative binomial distribution has a dispersion index of $d = 1 + \lambda/\nu = 1 + 1/\phi$ = (1.0, 1.5, 2.0, 3.0, 4.0)). To be clear, we consider:

- scenarios with $\phi = \infty$ and $\omega = 0$ as those with data simulated from the Poisson distribution;
- scenarios with $\phi < \infty$ and $\omega = 0$ as those with data simulated from the negative binomial distribution;
- scenarios with $\phi = \infty$ and $\omega > 0$ as those with data simulated from the zero-inflated Poisson distribution; and
- scenarios with $\phi < \infty$ and $\omega > 0$ as those with data simulated from the zero-inflated negative binomial distribution.

We simulated X_i as a univariate continuous covariate from a Normal distribution, with mean of zero and variance of 100: $X_i \sim \text{Normal}(0, 100)$, for i in 1, ..., n (as such, $k = 2$). Note that the covariate matrix X is simulated anew for each individual simulation run. Therefore, we are considering the case of *random* regressors. Chen and Giles (2011) discuss the difference between fixed and random covariates. The assumption of fixed covariates is generally considered only in experimental settings whereas an assumption of random covariates is typically more appropriate for observational studies.

Note that, for Poisson distributed data, we are simulating data with overall mean of $\lambda = \exp(\beta_0) \approx$ (1.6, 2.7, 4.5, 7.4, 12.2).

For $\lambda > 5$, zeros in the data are quite rare since $\Pr(Y = 0|\lambda) \approx 0$. The simulation study could be expanded in several ways. For instance, we did not consider models that deal with under-dispersion, even though under-dispersed counts may arise in various ecological studies; see Lynch et al. (2014). Also note that the simulation study only tests for rates of false positives (since $\beta_X = 0$ and $\gamma_X = 0$ for all scenarios). We are not testing for excessive false negatives (and overly wide confidence intervals) which are also undesirable.

In total, we considered 1,000 distinct scenarios and for each simulated 15,000 unique datasets. For each dataset, we conducted the seven-step procedure and recorded all *p*-values and whether or not the null hypotheses is rejected at the 0.05 significance level under the entire procedure. We also recorded all AIC, AICc and BIC statistics. We are interested in the unconditional false positive rate (i.e. the unconditional type 1 error rate).

We specifically chose to conduct 15,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a negligible amount (for looking at a false positive rate of $\alpha = 0.05$, Monte Carlo SE will be approximately $0.0018 \approx \sqrt{0.05\,(1-0.05)\,/15,000}$; see Morris et al., 2019). We ran all simulations using R on parallel nodes of the Compute Canada cluster; see Baldwin (2012).

To test the association between *X* and *Y* with each of the regression models, we conducted a Wald test (using the Chi-square statistic) to obtain the necessary *p*-value since in R, the *p*-values in the default summary.glm output are from Wald tests (using the Chi-square statistic). Moreover, in initial simulations, LRTs performed rather erratically in rare situations when the model was misspecified (e.g. when a Poisson model was fit to ZIP data). The glm function ('stats' package) was used to fit Poisson GLMs; the glm.nb function ('MASS' package) was used to fit the NB GLMs; and the zeroinfl function ('pscl' package) was used to fit the ZIP and ZINB GLMs; see R-code in Appendix S1.

3 | DISCUSSION

Analysis under the 'correct model'—We first wish to confirm that the models under investigation deliver correct type 1 error when used as intended. In other words, suppose the 'correct model' is known a priori and is used regardless of any preliminary diagnostic testing, would we obtain the desired number of false positives? See Figure 2 which plots the rejection rates corresponding to this question.

In summary, we see that for data simulated from a Poisson distribution (Figure 2, panel 1), empirical type 1 error is indeed

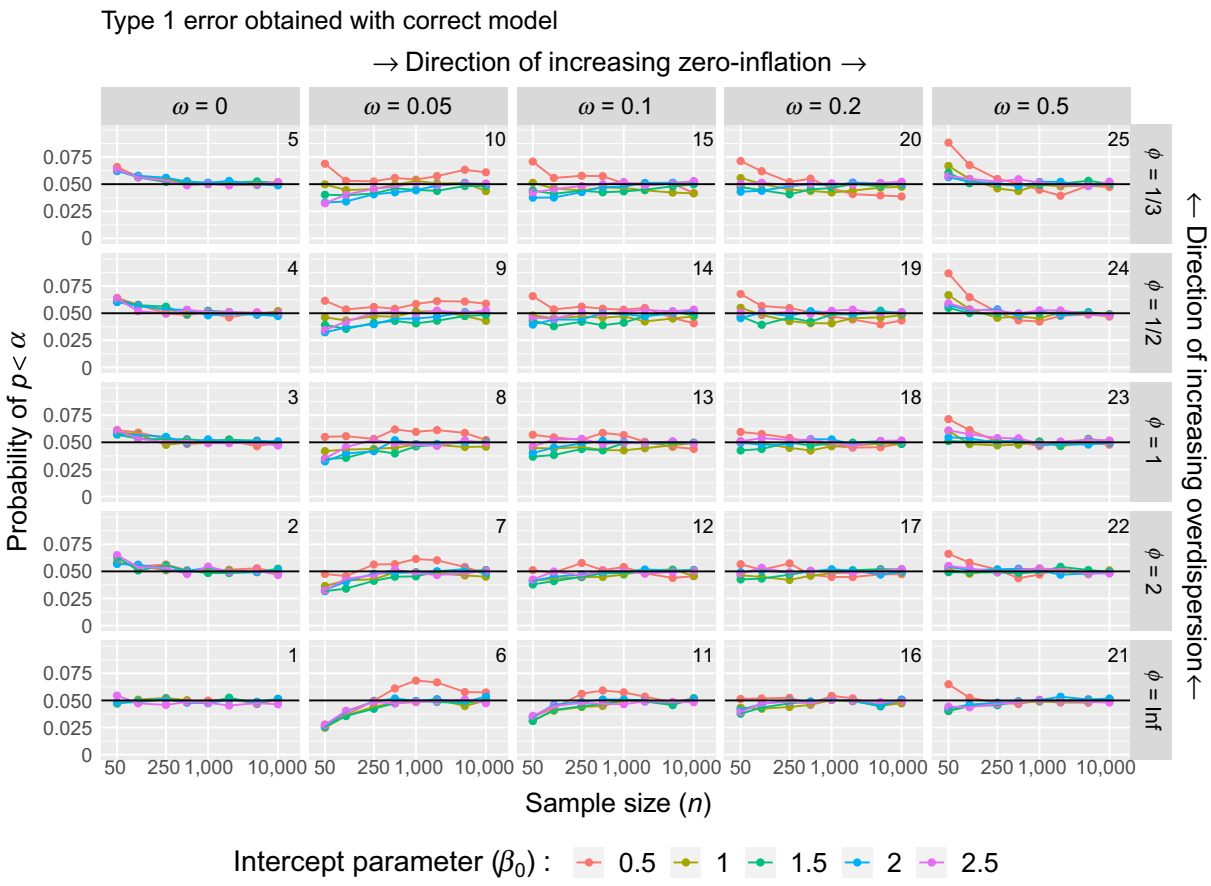

FIGURE 2 The empirical level of type 1 error (for testing the association between *X* and *Y*) obtained under the 'correct' model. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

approximately 0.05 as desired. We also note that for data simulated from the NB distribution (Figure 2, panels 2–5), empirical type 1 error is approximately 0.05 for all $n \geq 250$ and for all values of the intercept parameter (β_0). For data simulated from the ZIP distribution (see Figure 2, panels 6, 11, 16 and 21), empirical type 1 error can be substantially conservative (i.e. <0.05) for small sample sizes and when the probability of structural zeros is low (i.e. for small values of n and small values of ω). Finally, for ZINB data, we note that, when n is small, the false positive rate is higher than the advertised rate of 0.05 for some scenarios and less than 0.05 for others. For example, with $n = 100$, $\beta_0 = 0.5$, $\phi = 1/3$, and $\omega = 0.5$, the false positive rate is 0.07, whereas, when $n = 100$, $\beta_0 = 2.5$, $\phi = 2$, and $\omega = 0.05$, the type 1 error is 0.04 (see Figure 2, panels 25 and 7 respectively).

None of the models appear to be 'robust' to model misspecification. A Poisson model applied to non-Poisson data leads to very high rejection rates (so high they are often off the charts in Appendix S1—Figure 19). A ZIP model also performs poorly when applied to non-ZIP data (see Appendix S1—Figure 20), as does the NB model when applied to non-NB data (see Appendix S1—Figure 21), and the ZINB model (see Appendix S1—Figure 22) when applied to non-ZINB data (specifically when applied to Poisson data and NB data).

More specifically, it seems inadvisable to recommend simply fitting a ZIP or ZINB to Poisson data if one is uncertain about the possibility of zero-inflation, or overdispersion. As the sample size, n, increases (and as β_0 decreases), the type 1 error rates obtained when the ZIP and ZINB models are fit to Poisson data increase well beyond 0.05 (see Appendix S1—Figures 20 and 22, panel 1). This unexpected result may be due to the fact that these models are testing a null hypothesis that lies on the boundary of the parameter space (i.e. $\omega = 0$); see Feng and McCulloch (1992). In contrast, the NB model, when fit to Poisson data, maintains correct type 1 error for both small and large sample sizes. However, when fit to ZIP or ZINB data, the NB model results in either far too few or far too many false positives (depending on n and β_0); see Appendix S1—Figure 21.

Preliminary testing—The next question is: how often do the preliminary tests reject their null hypotheses? We also wish to determine how often the preliminary testing scheme successfully identifies the 'correct' model.

Let us first consider the D&L score test (see Appendix S1—Figure 16) and specifically as it applies to the NB scenarios. Recall that the NB scenarios are those with overdispersion ($\phi < \infty$) but no structural zero-inflation ($\omega = 0$). With the exception of the small sample-size scenarios with a small amount of overdispersion ($n \leq 100$, $\phi \leq 2$), the D&L test correctly rejects the null hypothesis of no overdispersion for the vast majority of cases (Appendix S1—Figure 16, panels 2–5). For Poisson data (when $\phi = \infty$ and $\omega = 0$), the D&L test shows approximately correct type 1 error, with rejection rates ranging from 0.039 to 0.056 (see Appendix S1—Figure 16, panel 1). However, for ZIP data (when $\phi = \infty$ and $\omega > 0$), the D&L test will often reject the null hypothesis of no overdispersion; see Appendix S1—Figure 16, panels 6, 11, 16 and 21. The rate of rejection increases with increasing sample size, with increasing ω, and with increasing β_0. Strictly speaking, rejection in these cases is correct since an excess of zeros ($\omega > 0$) does contribute to overdispersion. However, it must be noted that using the NB model for overdispersion when the underlying issue is zero-inflation is not appropriate; see Harrison (2014). Indeed, when the NB model is fit to ZIP data, we record type 1 error rates either much too low or much too high, depending on ω, β_0, and n; see Appendix S1—Figure 21, panels 6, 11, 16 and 21.

Now let us consider the Vuong test for zero-inflation. See Appendix S1—Figures 17 and 18 for the Vuong test results. Note that the 'Poisson versus ZIP' Vuong test will often reject the null of no zero-inflation for NB data (Appendix S1—Figure 17, panels 2–5). In contrast, the 'NB versus ZINB' Vuong test will rarely reject the null of no zero-inflation for NB data (Appendix S1—Figure 18, panels 2–5). In this way, the Vuong test acts as a second-line defense against erroneously selecting a Poisson model. If the D&L score test fails to select the NB model in Step 1, the 'Poisson versus ZIP' Vuong test in Step 3 will often reject a Poisson model in favour of the ZIP model (particularly when n and β_0 are large). The ZIP model, when used for NB, is not ideal, but is definitely preferable to a Poisson model; compare Appendix S1—Figures 19 and 20, panels 2–5.

Overall, the probability that the preliminary seven-step testing scheme selects the 'correct' model depends highly on β_0, ω, ϕ and n, see Figure 3. With Poisson data, if each of the diagnostic tests were truly independent (and each had a $\alpha = 0.05$ type 1 error rate), then the probability of selecting the 'correct' model should be 90.25% (=0.95 ×95%); see Figure 1. The numbers we obtain from the simulation study range from 87% to 96%.

For the ZIP data scenarios, the 'incorrect' ZINB model is chosen in a majority of cases ($\omega > 0$, $\phi = \infty$; Figure 3, panels 6, 11, 16 and 21). This may not necessarily lead to type 1 error inflation since the 'incorrect' ZINB model is often conservative when applied to ZIP data; see Appendix S1—Figure 22. For ZINB data scenarios (i.e. when $\omega > 0$, $\phi < \infty$), in cases when the ZINB model is not selected, it is most likely that the NB model is selected instead. This also might not necessarily lead to type 1 error inflation since the misspecified NB model appears to maintain a type 1 error rate at or bellow the advertised rate in many of these situations (specifically when $\phi < 2$ and $\omega < 0.2$); see Appendix S1—Figure 21.

Post-score-testing unconditional type 1 error—Our main question of interest is whether or not the null hypotheses of no association between X and Y is rejected at the desired 0.05 significance level when following the entire seven-step procedure outlined in Section 2.2. The corresponding rejection rates are plotted in Figure 4. Table 1 lists rejection rates and model selection rates for a select number of scenarios. Let us consider the results for each distribution.

First, for data simulated from the Poisson distribution (Figure 4, panel 1), empirical type 1 error appears to be unaffected by model selection bias. This is due to the fact that incorrect models are rarely selected, even when sample sizes are small (see Figure 3, panel 1). Consider two specific scenarios:

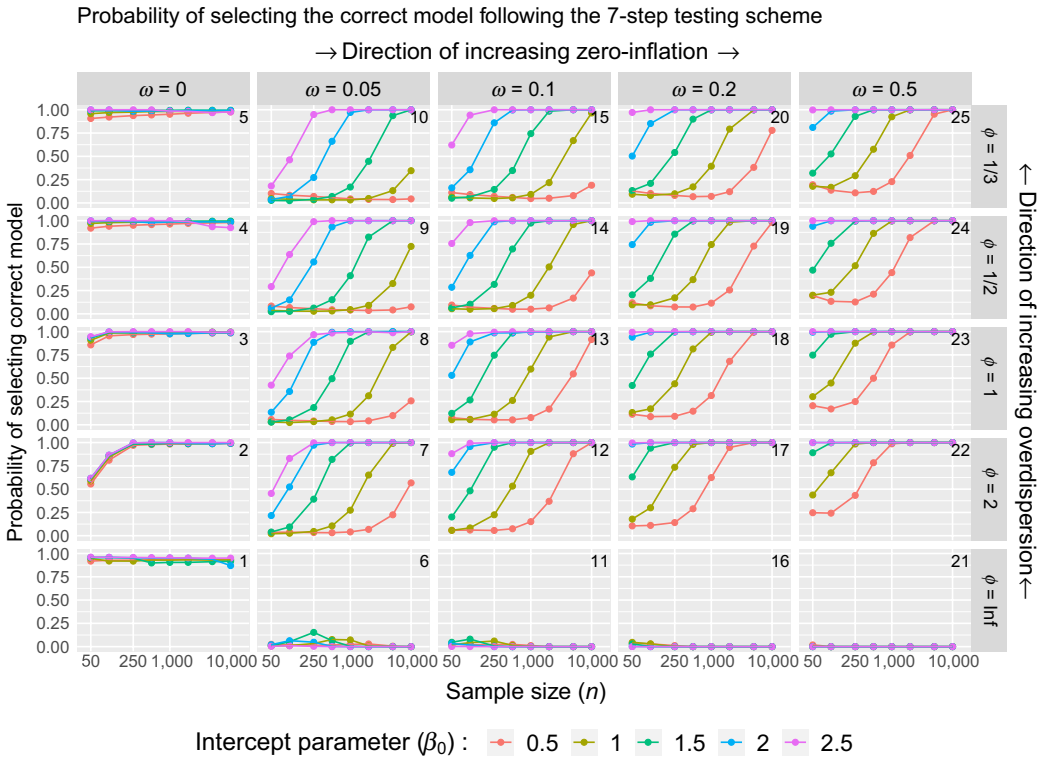

FIGURE 3 The probability of selecting the 'correct model' after following the seven step testing scheme outlined in Section 2.2. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

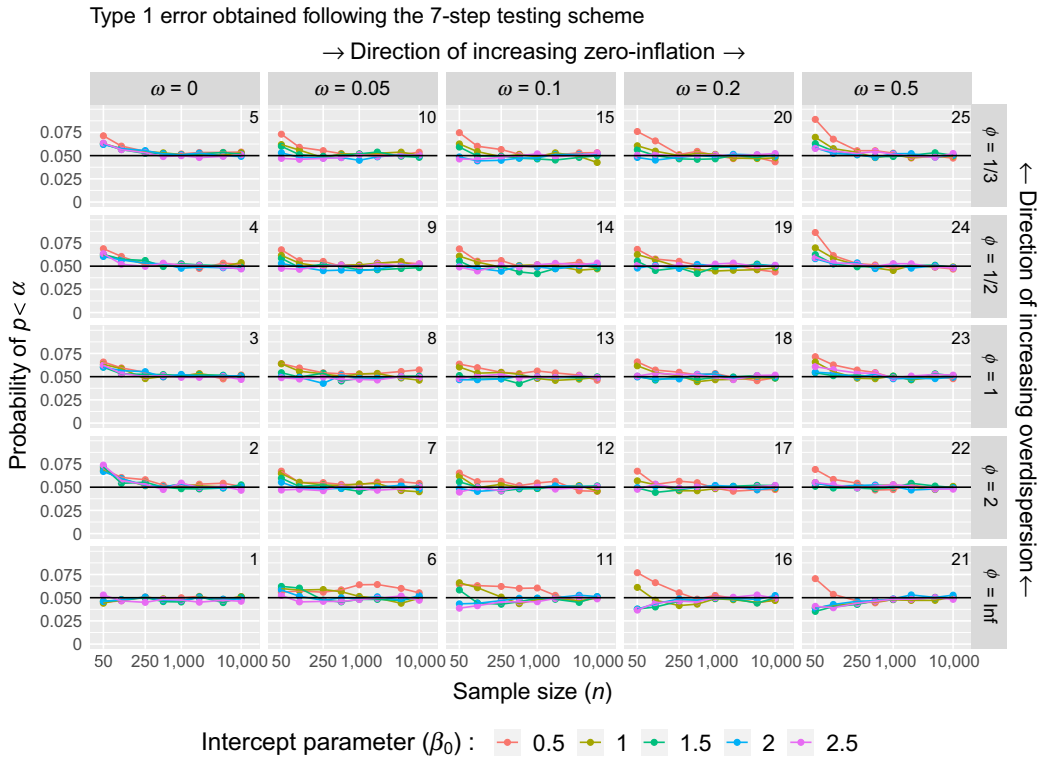

FIGURE 4 Type 1 error (for testing the association between *X* and *Y*) obtained following the seven step testing scheme outlined in Section 2.2. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

TABLE 1 Rejection rates and model selection rates for a number of selected scenarios from the simulation study; see coressponding Figures 9–13 in the Appendix S1. These numbers can be used to calculate the overall unconditional type 1 error rates. For example, for Scenario '402', the type 1 error obtained after model selection via AIC is 0.070 (=0.07 × 0.51 + 0.09 × 0.32 + 0.03 × 0.16 + 0.08 × 0.01); the type 1 error obtained after model selection via BIC is 0.073 (=0.07 × 0.84 + 0.20 × 0.06 + 0.02 × 0.10 + 0.20 × 0.00); and the type 1 error obtained after model selection via sequential score tests is 0.063 (=0.07 × 0.71 + 0.14 × 0.02 + 0.03 × 0.24 + 0.13 × 0.02)

	Poisson GLM	ZIP GLM	NB GLM	ZINB GLM
Scenario '3' ($n = 250$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0$; Poisson)				
Pr(reject H_0)	0.05	0.04	0.05	0.04
Pr(reject H_0|M selected by tests)	0.05	0.04	0.03	0.09
Pr(M selected by tests)	0.93	0.03	0.04	0.00
Pr(reject H_0|M has lowest AIC)	0.05	0.17	0.04	0.06
Pr(M has lowest AIC)	0.84	0.11	0.05	0.00
Pr(reject H_0|M has lowest BIC)	0.05	0.34	0.01	—
Pr(M has lowest BIC)	0.99	0.00	0.01	0.00
Scenario '6' ($n = 2{,}000$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0$; Poisson)				
Pr(reject H_0)	0.05	0.07	0.05	0.07
Pr(reject H_0|M selected by tests)	0.05	0.07	0.05	0.23
Pr(M selected by tests)	0.93	0.02	0.05	0.00
Pr(reject H_0|M has lowest AIC)	0.05	0.31	0.04	0.16
Pr(M has lowest AIC)	0.84	0.10	0.06	0.00
Pr(reject H_0|M has lowest BIC)	0.05	0.80	0.02	—
Pr(M has lowest BIC)	1.00	0.00	0.00	0.00
Scenario '46' ($n = 2{,}000$, $\beta_0 = 0.5$, $\phi = 2$, $\omega = 0$; NB)				
Pr(reject H_0)	0.11	0.08	0.05	0.07
Pr(reject H_0|M selected by tests)	—	—	0.05	0.17
Pr(M selected by tests)	0.00	0.00	0.99	0.01
Pr(reject H_0|M has lowest AIC)	—	—	0.05	0.25
Pr(M has lowest AIC)	0.00	0.00	0.87	0.13
Pr(reject H_0|M has lowest BIC)	—	—	0.05	0.17
Pr(M has lowest BIC)	0.00	0.00	1.00	0.00
Scenario '57' ($n = 50$, $\beta_0 = 1.5$, $\phi = 2$, $\omega = 0$; NB)				
Pr(reject H_0)	0.11	0.04	0.06	0.02
Pr(reject H_0|M selected by tests)	0.11	0.00	0.05	0.11
Pr(M selected by tests)	0.39	0.00	0.61	0.00
Pr(reject H_0|M has lowest AIC)	0.11	0.09	0.05	0.03
Pr(M has lowest AIC)	0.32	0.08	0.56	0.04
Pr(reject H_0|M has lowest BIC)	0.12	0.09	0.04	0.10
Pr(M has lowest BIC)	0.53	0.03	0.43	0.00
Scenario '251' ($n = 250$, $\beta_0 = 1$, $\phi = 2$, $\omega = 0.05$; ZINB)				
Pr(reject H_0)	0.12	0.08	0.05	0.04
Pr(reject H_0|M selected by tests)	0.00	—	0.05	0.13
Pr(M selected by tests)	0.00	0.00	0.95	0.05
Pr(reject H_0|M has lowest AIC)	0.00	0.09	0.05	0.08
Pr(M has lowest AIC)	0.00	0.04	0.63	0.33
Pr(reject H_0|M has lowest BIC)	0.00	0.14	0.05	0.23
Pr(M has lowest BIC)	0.00	0.05	0.93	0.01
Scenario '402' ($n = 100$, $\beta_0 = 0.5$, $\phi = \infty$, $\omega = 0.1$; ZIP)				
Pr(reject H_0)	0.07	0.05	0.05	0.04
Pr(reject H_0|M selected by tests)	0.07	0.14	0.03	0.12
Pr(M selected by tests)	0.71	0.02	0.24	0.02
Pr(reject H_0|M has lowest AIC)	0.07	0.09	0.03	0.08
Pr(M has lowest AIC)	0.51	0.32	0.16	0.01
Pr(reject H_0|M has lowest BIC)	0.07	0.20	0.02	0.20
Pr(M has lowest BIC)	0.84	0.06	0.10	0.00

- Scenario '3' ($n = 250$, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$)—When $\beta_0 = 0.5$ and $n = 250$, a Poisson model is correctly selected in approximately 93% of cases while the NB and ZIP models are selected in about 4% and 3% of cases, respectively. Numbers in the top right-hand corner of each node in Figure 1 indicate the expected number of datasets (out of a total of 100) to reach each outcome if the data were Poisson (with $\beta_X = 0$), and each of the tests were truly independent (with a $\alpha = 0.05$ type 1 error rate). The numbers in parentheses correspond to results from the simulation study for this scenario.

- Scenario '6' (n = 2,000, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$)—When $\beta_0 = 0.5$ and n = 2,000, a Poisson model is correctly selected in approximately 93% of cases while the NB and ZIP models are selected in about 5% and 2% of cases, respectively. While the NB model is perhaps conservative for this data (Pr(reject H_0|NB model selected by tests) = 0.049), the ZIP model is not (Pr(reject H_0|ZIP model selected by tests) = 0.070). However, the impact is negligible: the unconditional type 1 error rate obtained after following the seven-step procedure is 0.051.

Second, for data simulated from the ZIP distribution (i.e. when $\omega > 0$ and $\phi = \infty$), the 'incorrect' ZINB model is almost always selected due to the fact that the model selection procedure tests for zero-inflation only after first testing for overdispersion. However, the type 1 error under this 'incorrect' ZINB model is, for most scenarios, not substantially higher than the advertised 0.05 rate, (see Appendix S1—Figure 22, panels 6, 11, 16 and 21). There are, however, exceptions where model selection bias is apparent. Consider, for example, scenario '402':

- Scenario '402' (n = 100, $\beta_0 = 0.5$, $\phi = \infty$ and $\omega = 0.1$)—The unconditional type 1 error obtained after following the seven-step procedure is 0.063 (see Figure 4, panel 11). Among the simulated datasets for which the ZIP model is selected (by the D&L and Vuong tests), the ZIP model has a rejection rate of 0.14. Among the simulated datasets for which the ZINB model is selected, the ZINB model has a rejection rate of 0.12; see Table 1. This clearly shows that the diagnostic tests (the D&L and Vuong tests) and the subsequent hypothesis tests ($H_0 : \beta_X = \gamma_X = 0$) are not independent of one another. In this instance, the D&L test will not only screen for overdispersion, but will also direct the data towards a model that is more likely to reject $H_0 : \beta_X = \gamma_X = 0$, thereby inflating the type 1 error.

With data simulated from the NB distribution (i.e. when $\phi < \infty$ and $\omega = 0$; see Figure 4, panels 2–5), we see that model selection bias can lead to modest type 1 error inflation when n is small. When sample sizes are sufficiently large, there is little evidence of any substantial type 1 error inflation caused by model selection bias. Consider for example 'Scenario 57':

- Scenario '57' (n = 50, $\beta_0 = 1.5$, $\phi = 2$ and $\omega = 0$)—The unconditional type 1 error obtained after following the seven-step procedure is 0.072 (see Figure 4, panel 2), whereas the type 1 error obtained with the 'correct' NB model is 0.063. This inflation is due to the fact that, for these data, there is a 39% probability of selecting a Poisson model following the seven-step procedure and that Pr(reject H_0|Poisson model is selected by tests) = 0.114; see Table 1.

Finally, consider data simulated from the ZINB distribution (i.e. when $\phi < \infty$ and $\omega > 0$; see Figure 4, panels 7–10, 12–15, 17–20 and 22–25). We see type 1 error rates higher than 0.05 for small sample-size scenarios, but rates of approximately 0.05 otherwise. For example, consider scenario '251':

- Scenario '251' (n = 250, $\beta_0 = 1.0$, $\phi = 2$ and $\omega = 0.05$)—The unconditional type 1 error obtained after following the seven step procedure is 0.053 (see Figure 4, panel 7). The seven-step procedure correctly selects the ZINB GLM with a probability of only 4.6%. However, among the 95.4% of simulated datasets for which the NB model is incorrectly selected, the null hypothesis ($H_0 : \beta_X = \gamma_X = 0$) is rejected with probability of 0.048.

AIC, AICc and BIC model selection—We also investigated model selection using information criteria metrics. We were particularly curious as to how often the 'correct' model is the model with the lowest AIC/BIC; Figures 5 and 6 plot the results. The results for AICc are almost identical to those obtained for AIC; see Figure 14 in the Appendix S1.

We see that the probability that AIC selects the 'correct' model depends highly on β_0, ω, ϕ and n. Perhaps unexpectedly, we see that this probability does not necessarily increase with increasing sample size for certain scenarios. (The same is observed for AICc.) In contrast, the probability that the BIC selects the 'correct' model does increase with increasing sample size. Overall, averaging across all 1,000 scenarios we considered, AIC selected the correct model for 82% of datasets, AICc selected the correct model for 81% of datasets, the BIC selected the correct model for 73% of datasets, and the seven-step model selection procedure based on score tests selected the correct model for 63% of datasets.

We also wish to determine whether or not the null hypotheses of no association between X and Y is rejected at the 0.05 significance level when following model selection via AIC/AICc/BIC. Figure 7 shows that, when β_0 is small, there are several scenarios in which the post-AIC unconditional type 1 error is much higher than 0.05. (The same is observed with AICc, see Figure 15 in the Appendix S1.) Perhaps most surprisingly, we see that, with Poisson data (i.e. scenarios with $\omega = 0$ and $\phi = \infty$), the post-AIC (post-AICc) unconditional type 1 error increases with increasing sample size (e.g. with $\beta_0 = 0.5$, the post-AIC (post-AICc) unconditional type 1 error increases from 0.050 to 0.085 (from 0.501 to 0.085) as n increases from 50 to 10,000; see Figure 7 (see Appendix S1—Figure 15), panel 1). This does not appear to be an issue for the BIC (Figure 8).

Consider again Scenario '3' (n = 250, $\beta_0 = 0.5$, $\phi = \infty$, and $\omega = 0$) and Scenario '6' (n = 2,000, $\beta_0 = 0.5$, $\phi = \infty$ and $\omega = 0$); see Table 1:

- For Scenario '3', the probability that AIC correctly selects the Poisson GLM is high at 84%. However, there is a 11% probability that the ZIP GLM is selected and Pr(reject H_0|ZIP GLM has lowest AIC) = 0.17. This drives the unconditional type 1 error rate to 0.062 (=0.05 × 0.84 + 0.17 × 0.11 + 0.04 × 0.05 + 0.06 × 0.00). (The unconditional type 1 error rate obtained with the AICc is also 0.062). The BIC correctly selects the Poisson GLM with probability of 99%. While we have that Pr(reject H_0|ZIP GLM has lowest BIC) = 0.50, there is a less than 1% probability that the ZIP GLM is selected. As such, the post-BIC unconditional type 1 error rate remains low at 0.053 (=0.05 × 0.99 + 0.34 × 0.00 + 0.01 × 0.01).
- For Scenario '6', the probability that AIC incorrectly selects the ZIP GLM is 0.10, and Pr(reject H_0|ZIP GLM has lowest AIC) = 0.31.

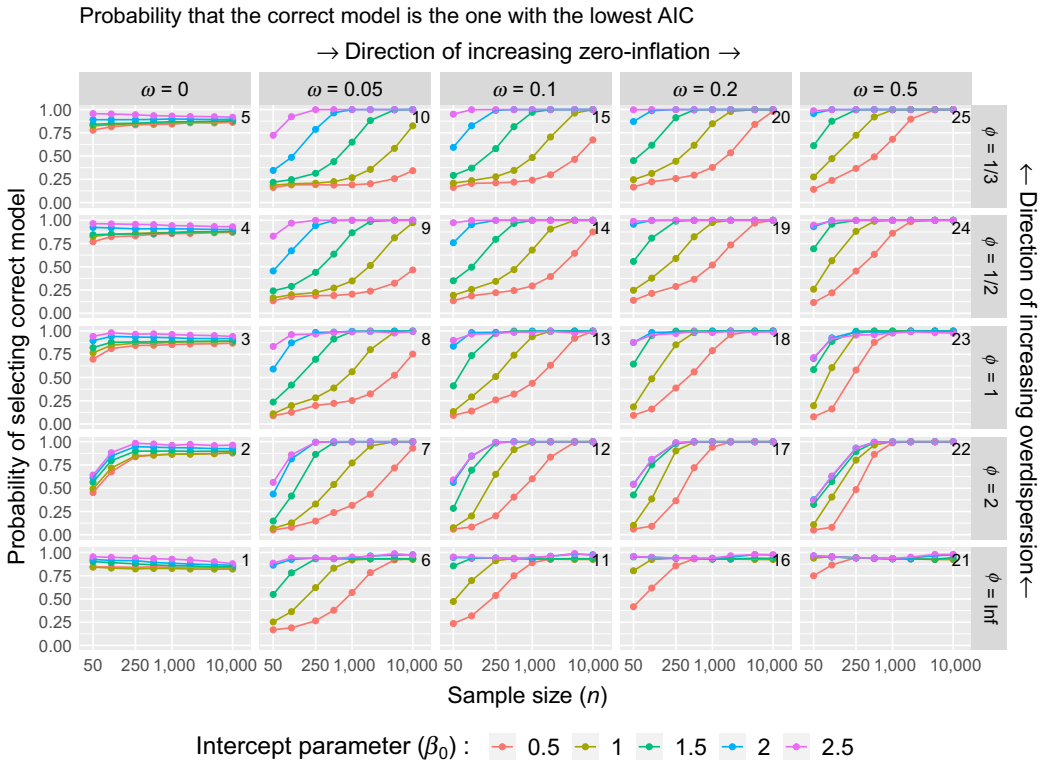

FIGURE 5 The probability that the 'correct' model is the one with the lowest AIC. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

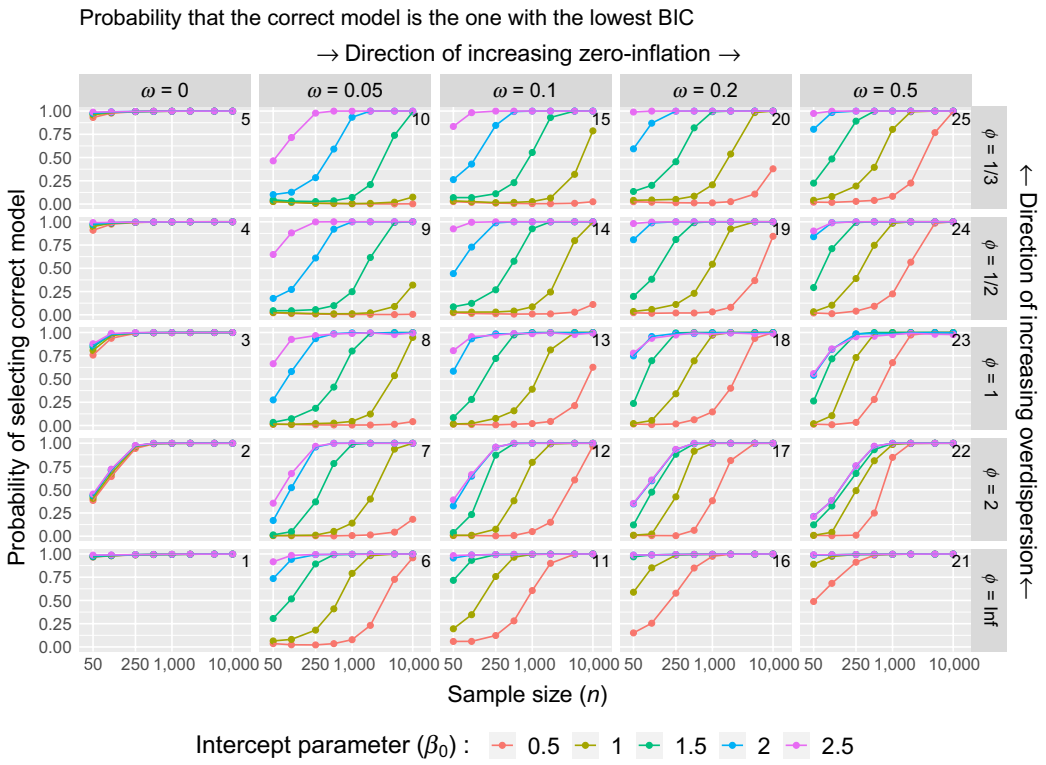

FIGURE 6 The probability that the 'correct' model is the one with the lowest BIC. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

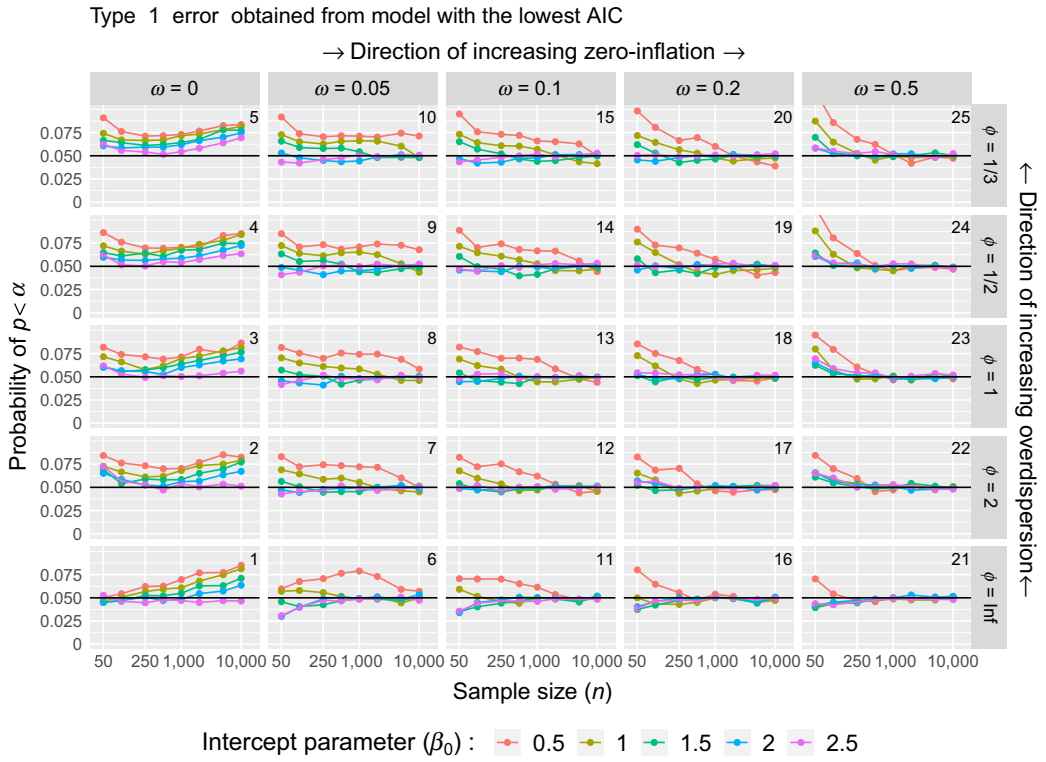

FIGURE 7 Type 1 error (for testing the association between *X* and *Y*) obtained from model with the lowest AIC. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

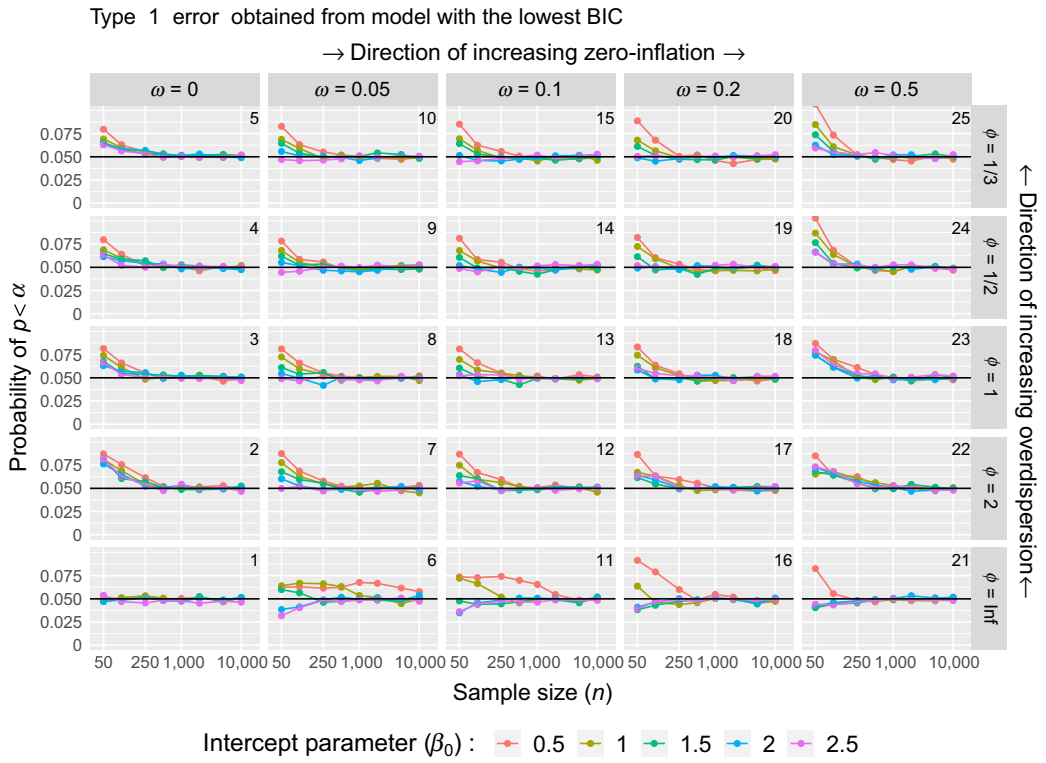

FIGURE 8 Type 1 error (for testing the association between *X* and *Y*) obtained from model with the lowest BIC. For panel 1, the 'correct' model is a Poisson GLM; for panels 2–5 the 'correct' model is the NB GLM; for panels 6, 11, 16 and 21, the 'correct' model is the ZIP GLM; and for other panels, the 'correct' model is the ZINB GLM

As such, the post-AIC unconditional type 1 error rate = 0.077 (=0.05 × 0.84 + 0.31 × 0.10 + 0.04 × 0.06 + 0.16 × 0.00). (The unconditional type 1 error rate obtained with the AICc is also 0.077). In contrast, the probability that the BIC correctly selects the Poisson GLM is more than 99% and the post-BIC unconditional type 1 error rate = 0.051.

With NB data (i.e. scenarios with $\omega = 0$ and $\phi < \infty$), unconditional type 1 error rates can also be much higher than 0.05, even when n and β_0 are large. This is due to the fact that the ZINB model, when erroneously selected in a minority of cases, rejects the null of no association between X and Y at rates much much higher than 0.05 (particularly when n is large). Consider for example, Scenario '46':

- Scenario '46' ($n = 2{,}000$, $\beta_0 = 0.5$, $\phi = 2$ and $\omega = 0$)—Among the 87% of datasets for which AIC correctly selects the NB model, the null hypothesis of no association between X and Y is rejected with probability of exactly 0.050; see Table 1. However, among the remaining 13% of datasets for which the ZINB model is erroneously selected, the probability of rejecting the null hypothesis of no association between X and Y is 0.25. As a result the post-AIC unconditional type 1 error rate is 0.077 (=0.25 × 0.13 + 0.05 × 0.87). (The unconditional type 1 error rate obtained with the AICc is also 0.077). In contrast, the BIC correctly selects the NB model with probability of more than 99% and the post-BIC unconditional type 1 error rate = 0.052.

With ZIP and ZINB data, the BIC is often less capable of selecting the 'correct' model relative to the AIC and AICc. However, the BIC is still often preferable to the AIC and AICc in terms of maintaining the desired unconditional type 1 error rate. Consider for example Scenario '251' ($n = 250$, $\beta_0 = 1.0$, $\phi = 2$ and $\omega = 0.05$); see Table 1:

- For Scenario '251', the probability that AIC (AICc) correctly selects the ZINB GLM is 32% (31%) and the probability that the ZINB GLM will reject the null ($\beta_X = \gamma_X = 0$) for these data is 0.043. However, these are not independent events. Indeed, we have Pr(reject H_0|ZINB GLM has lowest AIC) = 0.079 which increases the unconditional type 1 error rate to 0.059. (The post-AICc unconditional type 1 error rate is also 0.059). The BIC correctly selects the ZINB GLM with probability of only 1%. With a probability of 93%, the NB GLM is selected instead. However, since Pr(reject H_0|NB GLM has lowest BIC) = 0.05, the post-BIC unconditional type 1 error rate remains relatively low at only 0.055. We can compare this to the unconditional type 1 error rate obtained following the seven-step score testing procedure of 0.053.

In summary, while AIC (or AICc) is often able to select the 'correct' model more frequently than the BIC or the sequential score testing scheme, there may be a greater potential for type 1 error inflation. Indeed, averaging across all 1,000 unique scenarios considered, the post-AIC unconditional type 1 error obtained is 0.055 (the post-AICc unconditional type 1 error obtained is 0.056), whereas the post-BIC unconditional type 1 error obtained is 0.053, and the post-score testing unconditional type 1 error obtained is 0.051; see Table 2. How can this be?

TABLE 2 For each of the four different model selection schemes, the unconditional type 1 error and the probability of selecting the correct model are averaged across all 1,000 simulation scenarios

	Type 1 error	Prob. of selecting the correct model
Seven-step procedure	5.13%	63.06%
AIC-based model selection	5.55%	81.51%
AICc-based model selection	5.56%	80.90%
BIC-based model selection	5.29%	73.03%

In the presence of model selection bias, selecting the 'correct' model more often is, somewhat paradoxically, not necessarily preferable. This is due to the fact that the model selection procedure based on AIC/AICc/BIC and the hypothesis test for the association between X and Y are clearly not independent. The seven-step testing scheme comes out more favourably in our simulation study for the simple reason that, in the first step, the test for overdispersion often leads one to select the relatively robust (yet not necessarily correct) NB model. One might therefore reasonably conclude that the seven-step procedure gives better results, but for the wrong reasons.

4 | CONCLUSIONS

'Model misspecification is a major, if not the dominant, source of error in the quantification of most scientific evidence', writes Taper (2004). With this in mind, it is no surprise that researchers are repeatedly advised to do whatever is necessary so as to avoid fitting their data with a misspecified model. However, post hoc model selection can come with unintended consequences. Indeed, researchers, in their sincere efforts to select the 'best model', should not forget the potential for the collateral damage that is model selection bias.

If the population distribution is known in advance, model selection bias will not be a problem. If the assumptions required of the Poisson distribution are known to be wrong, alternative models that do not depend on these assumptions can be used and ideally a valid model can be pre-specified prior to obtaining/observing any data. However, outside of a highly controlled laboratory experiment, this may not be realistic. The potentially problematic (and most likely scenario) is when one cannot, with a high degree of confidence, determine the distributional nature of the data before observing the data. What should be done in these circumstances? Tsou (2006) suggest using a 'robust' Poisson regression model 'so that one need not worry about the correctness of the Poisson assumption'. However, when the distributional assumptions of a Poisson GLM do hold, Tsou (2006) acknowledge that the 'robust approach might not be as efficient'. Given the potentially immense expense required to obtain data, anyone working in data-driven research will no doubt

be reluctant to adopt any approach which compromises statistical power.

Researchers who do not know in advance whether or not there is overdispersion or zero-inflation, might decide to simply use a ZIP or ZINB as a 'safer bet' (Perumean-Chaney et al., 2013) and pay a price in terms of efficiency (Williamson et al., 2007). However, this is problematic. We observed that the ZIP and ZINB models, when fit to ordinary Poisson data, can lead to type 1 error well above the advertised rate when sample sizes are large. (Future work should consider whether hurdle models (Rose et al., 2006) are similarly problematic.) Instead, if there is sufficient data, researchers should proceed with a model selection procedure, ideally one based on efficient score tests. Our simulation study suggests that, if sample sizes are sufficiently large, there is little need to worry about model selection bias following a series of sequential score tests. However, when sample sizes are small, our simulation study demonstrated that model selection bias can lead to potentially substantial type 1 error inflation.

An interesting result from the simulation study is that AIC (or AICc) and BIC are likely to choose the correct model more often than the seven-step procedure. However, this does not mean that model selection based on AIC (or AICc) can be recommended. We observed that, even when sample sizes are large and when the true underlying distribution of the data is Poisson, using AIC (or AICc) to select the 'best' model can lead to substantial type 1 error inflation. The BIC is less problematic.

Future work should investigate the suitability of other model-selection criteria. These include simulation-based methods; see Brewer et al. (2016), Dunn and Smyth (1996) and Hartig (2017). While simulation is considered by some to be a powerful tool for assessing model fit, it remains 'rarely used' (Harrison et al., 2018). It would be beneficial to assess how simulation-based tests for overdispersion and zero-inflation compare to the more traditional score-based tests we considered in our study.

Regardless of the method used for model selection, the process of selecting a model for inference based on the data ignores a crucial source of uncertainty. And, if the model selection procedure and the hypothesis testing which follows are not independent, type 1 error inflation may occur. Ignoring the possibility of overdispersion and zero-inflation during data analyses can lead to invalid inference. However, if one does not have sufficient power to confidently test for overdispersion and zero-inflation, it may be best to simply use a model that can accommodate for these possibilities (e.g. use a robust model) instead of going through a model selection procedure that might inflate the type 1 error.

In summary, if one does not have the power to test for distributional assumptions, testing for distributional assumptions may not be wise. And if one does have a sufficiently large sample size to test for distributional assumptions, testing for distributional assumptions may be very beneficial. Note that our simulation study only included a single covariate and in studies where there are several covariates, it will no doubt be difficult to determine what constitutes a 'sufficiently large' sample size. To conclude, we note that researchers should always be cautious when interpreting results with smaller sample sizes (Button et al., 2013). Model selection bias is just one more reason to have a healthy skepticism of null hypothesis significance testing when sample sizes are small.

ACKNOWLEDGEMENTS

Many thanks to Marie Auger-Méthé and Paul Gustafson at the University of British Columbia for their input and support. Also, thank you to Natalie Pilakouta whose flowchart on how to analyse zero-inflated count data inspired much of this work and to the organizers of the 2020 International Statistical Ecology Conference.

DATA AVAILABILITY STATEMENT

All code used for simulation and to create plots is made publicly available at https://github.com/harlanhappydog/CountConsequences and on Zenodo at https://doi.org/10.5281/zenodo.4435386 (Campbell, 2020).

ORCID


Harlan Campbell https://orcid.org/0000-0002-0959-1594

REFERENCES

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Anderson, D. R. (2007). *Model based inference in the life sciences: A primer on evidence*. Springer Science & Business Media.

Baldwin, S. (2012). Compute canada: Advancing computational research. *Journal of Physics: Conference Series*, *341*, 012001. IOP Publishing.

Bening, V. E., & Korolev, V. Y. (2012). *Generalized Poisson models and their applications in insurance and finance*. Walter de Gruyter.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*(7), 949–959. https://doi.org/10.1111/2041-210X.13185

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, *87*(419), 738–754. https://doi.org/10.1080/01621459.1992.10475276

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603–618.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, *96*(9), 2370–2382.

Campbell, H. (2020). Data from: Files to accompany 'The consequences of checking for zero-inflation and overdispersion in the analysis of count data'. *Zenodo*, https://doi.org/10.5281/zenodo.4435386

Campbell, H., & Dean, C. (2014). The consequences of proportional hazards based model selection. *Statistics in Medicine*, *33*(6), 1042–1056.

Campbell, H., & Gustafson, P. (2019). The world of research has gone berserk: Modeling the consequences of requiring "greater statistical stringency" for scientific publication. *The American Statistician*, *73*(suppl 1), 358–373.

Chen, Q., & Giles, D. E. (2011). Finite-sample properties of the maximum likelihood estimator for the Poisson regression model with random covariates. *Communications in Statistics-Theory and Methods*, *40*(6), 1000–1014.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, *70*(1), 269–274.

Dean, C., & Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, *84*(406), 467–472.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv*. https://doi.org/10.1101/2020.04.26.048306

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Dushoff, J., Kain, M. P., & Bolker, B. M. (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*, *10*(6), 756–759. https://doi.org/10.1111/2041-210X.13159

Feng, Z., & McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, *13*(4), 325–332. https://doi.org/10.1016/0167-7152(92)90042-4

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics*, *6*(1), 17–24. https://doi.org/10.2307/3001420

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Freckleton, R. (2009). The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, *22*(7), 1367–1375. https://doi.org/10.1111/j.1420-9101.2009.01757.x

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University.

Greene, W. H. (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. NYU Working paper no. EC-94-10.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616. https://doi.org/10.7717/peerj.616

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794. https://doi.org/10.7717/peerj.4794

Hartig, F. (2017). *Dharma: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R package version 0.1 5(5).

Hilbe, J. M., & Greene, W. H. (2007). 7 count response regression models. *Handbook of Statistics*, *27*, 210–252.

Hooten, M. B., & Hefley, T. J. (2019). *Bringing Bayesian models to life*. CRC Press.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Hurvich, C. M., & Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44*(3), 214–217.

Kahan, B. C. (2013). Bias in randomised factorial trials. *Statistics in Medicine*, *32*(26), 4540–4549. https://doi.org/10.1002/sim.5869

Kelly, C. (2019). Rate and success of study replication in ecology and evolution. *PeerJ*, *7*(e7654). https://doi.org/10.7717/peerj.7654

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535. https://doi.org/10.1038/nn.2303

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.2307/1269547

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414–1421. https://doi.org/10.1890/10-1831.1

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

Lynch, H. J., Thorson, J. T., & Shelton, A. O. (2014). Dealing with under-and over-dispersed count data in life history, spatial, and community ecology. *Ecology*, *95*(11), 3173–3180. https://doi.org/10.1890/13-1912.1

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*(11), 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Murtaugh, P. A. (2014). In defense of p values. *Ecology*, *95*(3), 611–617.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: What's one to do? *Journal of Statistical Computation and Simulation*, *83*(9), 1671–1683. https://doi.org/10.1080/00949655.2012.668550

Potts, J. M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling*, *199*(2), 153–161. https://doi.org/10.1016/j.ecolmodel.2006.05.025

Puig, P., & Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, *101*(473), 332–340. https://doi.org/10.1198/016214505000000718

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, *45*(1), 218–227. https://doi.org/10.1111/j.1365-2664.2007.01377.x

Ridout, M., Hinde, J., & Demétrio, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, *57*(1), 219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, *12*(1), 81. https://doi.org/10.1186/1471-2288-12-81

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, *16*(4), 463–481. https://doi.org/10.1080/10543400600719384

Shuster, J. J. (2005). Diagnostics for assumptions in moderate to large simple clinical trials: Do they really help? *Statistics in Medicine*, *24*(16), 2431–2438. https://doi.org/10.1002/sim.2175

Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Martinez Del Rio, C. (2005). Information theory and hypothesis testing: A call for pluralism. *Journal of Applied Ecology*, *42*(1), 4–12. https://doi.org/10.1111/j.1365-2664.2005.01002.x

Taper, M. L. (2004). Model identification from many candidates. In M. L. Taper, & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 488–524). University of Chicago Press.

Tsou, T.-S. (2006). Robust Poisson regression. *Journal of Statistical Planning and Inference*, *136*(9), 3173–3186.

Uusipaikka, E. (2008). *Confidence intervals in generalized regression models*. Chapman and Hall/CRC.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, *51*(2), 738–743. https://doi.org/10.2307/2532959

Venzon, D., & Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *37*(1), 87–94.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*, 307–333.

Walters, G. D. (2007). Using Poisson class regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem. *Criminal Justice and Behavior*, *34*(12), 1659–1674. https://doi.org/10.1177/0093854807307030

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, *44*(5), 495–502. https://doi.org/10.1002/pits.20241

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

Whittingham, M. J., Swetnam, R. D., Wilson, J. D., Chamberlain, D. E., & Freckleton, R. P. (2005). Habitat selection by yellowhammers emberiza citrinella on lowland farmland at two spatial scales: Implications for conservation management. *Journal of Applied Ecology*, *42*(2), 270–280. https://doi.org/10.1111/j.1365-2664.2005.01007.x

Williams, M. N., & Albers, C. (2019). Dealing with distributional assumptions in preregistered research. *Meta-Psychology*, *3*, 1592.

Williamson, J. M., Lin, H., Lyles, R. H., & Hightower, A. W. (2007). Power calculations for ZIP and ZINB models. *Journal of Data Science*, *5*(4), 519–534.

Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE*, *10*(7), e0129606. https://doi.org/10.1371/journal.pone.0129606

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*(4), 937–950.

Yang, Z., Hardin, J. W., & Addy, C. L. (2010). Score tests for zero-inflation in overdispersed count data. *Communications in Statistics – Theory and Methods*, *39*(11), 2008–2030. https://doi.org/10.1080/03610920902948228

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1–25.

Zoltowski, D., & Pillow, J. W. (2018). Scaling the Poisson GLM to massive neural datasets through polynomial approximations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 3521–3531). Curran Associates, Inc.

Zorn, C. J. (1998). An analytic and empirical examination of zero-inflated and hurdle Poisson specifications. *Sociological Methods & Research*, *26*(3), 368–400. https://doi.org/10.1177/0049124198026003004

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Campbell H. The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods Ecol Evol.* 2021;00:1–16. https://doi.org/10.1111/2041-210X.13559