A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials

Journal:	*Biometrical Journal*
Manuscript ID	bimj.201700103.R1
Wiley - Manuscript type:	Research Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Zheng, Han; University of Southampton Faculty of Medicine, Primary Care and Population Sciences Kimber, Alan; University of Southampton, Mathematical Sciences Goodwin, Victoria; University of Exeter Medical School, NIHR CLAHRC South West Peninsula (PenCLAHRC) Pickering, Ruth; University of Southampton Faculty of Medicine, Primary Care and Population Sciences
Keywords:	Baseline counts, Negative binomial, Regression, Simulations

TITLE: A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials

RUNNING HEADER: Baseline counts in negative binomial models

AUTHORS: **Han Zheng** (hz9g13@soton.ac.uk), Medical Statistics Group, Faculty of Medicine, University of Southampton

Alan Kimber (A.C.Kimber@soton.ac.uk), Mathematical Sciences, University of Southampton

Victoria A Goodwin (v.goodwin@exeter.ac.uk), NIHR CLAHRC South West Peninsula (PenCLAHRC), University of Exeter Medical School

Ruth M Pickering (rmp@soton.ac.uk), Medical Statistics Group, Faculty of Medicine, University of Southampton

Address for correspondence
Dr RM Pickering, PhD CStat
Medical Statistics Group
Faculty of Medicine
Mailpoint 805, Level B
South Academic Block
Southampton General Hospital
SOUTHAMPTON, SO16 6YD

Tel: 023 80796565
Fax: 023 80796529
e-mail:rmp@soton.ac.u

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials

Han Zheng[1], Alan Kimber[2], Victoria A. Goodwin[3], and Ruth M. Pickering[*,1]

[1] Medical Statistics Group, Faculty of Medicine, University of Southampton
[2] Mathematical Sciences, University of Southampton
[3] NIHR CLAHRC South West Peninsula (PenCLAHRC), University of Exeter Medical School

Received zzz, revised zzz, accepted zzz

A common design for a falls prevention trial is to assess falling at baseline, randomize participants into an intervention or control group, and ask them to record the number of falls they experience during a follow-up period of time. This paper addresses how best to include the baseline count in the analysis of the follow-up count of falls in Negative Binomial (NB) regression. We examine the performance of various approaches in simulated datasets where both counts are generated from a mixed Poisson distribution with shared random subject effect. Including the baseline count after log-transformation as a regressor in NB regression (*NB-logged*) or as an offset (*NB-offset*) resulted in greater power than including the untransformed baseline count (*NB-unlogged*). Cook and Wei's Conditional Negative Binomial (*CNB*) model replicates the underlying process generating the data. In our motivating dataset, a statistically significant intervention effect resulted from the *NB-logged*, *NB-offset* and *CNB* models, but not from *NB-unlogged*, and large, outlying baseline counts were overly influential in *NB-unlogged* but not in *NB-logged*. We conclude that there is little to lose by including the log-transformed baseline count in standard NB regression compared to *CNB* for moderate to larger sized datasets.

Key words: Baseline counts; Negative binomial; Regression; Simulations

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX (please delete if not applicable)

1 Introduction

The goal of a falls prevention trial is to test whether an intervention is effective in reducing the occurrence of falls experienced by participants. A common design adopted in these trials is that, following an initial assessment of falling, participants who consent are randomized to either an intervention or control group, and they record an outcome number of falls during a follow-up period of time. A traditional and practical approach to analysis of the resulting outcome count has been to dichotomize and fit logistic regression, however information is lost during this process. An alternative analysis that is increasingly being used is to fit the outcome count in a regression for count responses (Donaldson *et al.*, 2009). Standard Poisson regression does not incorporate variability over participants, but this variability can be modelled as a random effect in a mixed Poisson distribution, and when assumed to follow a gamma distribution, leads to the Negative Binomial (NB) model (Hilbe, 2011). In both Poisson and NB regression the intervention effect is parameterized to yield a Falls Rate Ratio (FRR), that is, the ratio of the falls rate in the intervention group divided by that in the control group.

The use of the NB regression model to examine recurrent events was described for a medical audience by Glynn and Buring (1996), and the model has recently been recommended for the analysis of count

[*] Corresponding author: e-mail: rmp@soton.ac.uk, Phone: +44-238-079-6565, Fax: +44-238-079-6529

outcomes from falls prevention trials by the authors of the Cochrane Review (Gillespie *et al.*, 2012), but it is not clear how this is to be done in practice. In the description of statistical analysis performed in the trials included in the Cochrane review, the baseline count may be included as a discrete covariate, dichotomised or categorized, but details were not usually given.

Cook and Wei (2003) have proposed an alternative NB based model for recurrent event data with baseline counts. Response and baseline counts are assumed to follow a mixed Poisson distribution with a shared gamma distributed random subject effect, and the distribution of the response conditional on the observed baseline count is modelled. Unlike a standard NB regression, an estimate of the variance of the random subject effect is obtained using information from both the baseline and the follow-up counts. It is not uncommon for falls prevention trials to set a threshold on the baseline count as an eligibility criterion. In the Cook and Wei (2003) model, it is necessary to account for such a threshold in order to avoid the bias in the estimate of random subject variability that would otherwise occur.

Our analysis is motivated by experience of falls prevention trials in people with Parkinson's (PwP), and in particular the randomized controlled trial reported by Goodwin *et al.* (2011) in which PwP were recruited and randomized to an intervention group (receiving a 10-week strength and balance training programme) or a control group (receiving usual care). Participants recorded the follow-up falls they experienced prospectively using a daily diary during the 10-week intervention period. Prospective recording of falls using diaries is now the recommended method of collecting falls information (Hauer *et al.*, 2006). Goodwin *et al.* used the same method to collect a baseline count of falls during the 10-week period between recruitment and randomization: baseline and follow-up counts were available for 124 PwP (n=61 intervention; n=63 control). Participation in the trial was restricted to those who had experienced two or more falls in the previous year, but this was obtained from a retrospective single question at initial recruitment and was not a restriction on the prospectively recorded falls during the baseline period (which took value zero for some participants). It was also collected using different methodology and thus did not relate to the same process that produced the baseline or follow-up counts. Counts of falls typically follow a right-skewed distribution with a long tail: in the case of PwP extremely large numbers of falls can be recorded which may be influential in statistical analysis. This pattern can be seen in the scatterplot of baseline and follow-up counts from the trial reported by Goodwin *et al.* (Figure 1a), and after logarithmic transformation of both axes (Figure 1b). In isolation, the large counts might be considered outliers but seen in the context of the scatterplot they appear in keeping with a broadly linear relationship. The baseline count is likely to explain at least part of the random subject variability in a mixed Poisson distribution fitted to the outcome count, and it is also likely to be associated with unobserved prognostic factors. Although the CNB model accounts for the heterogeneity, it is not widely available to clinical researchers, which leads to a question – is it possible to fit an NB regression that accommodates the heterogeneity shared in the follow-up and baseline count following the underlying distribution of the CNB model?In this paper, we investigate various methods used in practice to incorporate the baseline count of falls as a discrete covariate in the analysis of the outcome count in standard NB regression, and compare these to Cook and Wei's (2003) CNB model as the benchmark. We fit models to the outcome count in our motivating dataset and examine the influence of large counts on model fit. NB regression is widely available in statistical packages, and the Wald test is typically the default option for assessing the significance of explanatory variables in the model. For this reason, the Wald test is the focus of our investigation, but we make a comparison to the performance of the score test. We simulate counts in scenarios reflecting our motivating dataset, and compare models with respect to the power and type I error rate of Wald tests, bias, and accuracy of model based standard error (SE) of the intervention effect.

2 Models for falls data

2.1 Mixed Poisson distribution

We assume that m participants ($i=1,2,\ldots,m$) are recruited to a trial, and prospectively count the falls y_{ij} (with time indicator $j=0$) for a baseline period. We also assume that the baseline phase lasts for the

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

same length of time (t_0) for all participants, which is common in falls prevention trials. At randomization, participants are allocated to an intervention (x_i=1) or control (x_i=0) group. They record y_{i1} falls during a follow-up period (with time indicator j=1) of length t_{i1}. Note that t_{i1} may differ across participants if they are lost to follow-up (assumed to occur at random) but in a trial t_{i1} would usually be planned to be the same for all participants. Variables Y_{i0} and Y_{i1} are non-negative integers, and if they are both Poisson distributed

$$\mathrm{P}(Y_{i0} = y_{i0}; \lambda_0, t_0) = \frac{(\lambda_0 t_0)^{y_{i0}} \exp(-\lambda_0 t_0)}{y_{i0}!} \tag{1}$$

$$\mathrm{P}(Y_{i1} = y_{i1}; \lambda_1, t_{i1}) = \frac{(\lambda_1 \exp(\beta x_i)\, t_{i1})^{y_{i1}} \exp(-\lambda_1 \exp(\beta x_i)\, t_{i1})}{y_{i1}!}, \tag{2}$$

where λ_0 is the average baseline falls rate and λ_1 is the average falls rate in the control group during the follow-up period. The FRR, the risk ratio $\mathrm{E}(Y_{i1}|x_i{=}1)/\mathrm{E}(Y_{i1}|x_i{=}0)$, is given by $\exp(\beta)$, where β is the parameter related to the intervention indicator x_i. Let $\mu_{i0}{=}\lambda_0 t_0$ at baseline and $\mu_{i1}{=}\lambda_1 \exp(\beta x_i)\, t_{i1}$ at follow-up. Then the expectation and variance of Y_{i0} and Y_{i1} are

$$\mathrm{E}(Y_{i0}) = \mathrm{Var}(Y_{i0}) = \mu_{i0} \tag{3}$$

$$\mathrm{E}(Y_{i1}) = \mathrm{Var}(Y_{i1}) = \mu_{i1}. \tag{4}$$

If baseline and follow-up counts are overdispersed, that is $\mathrm{Var}(Y_{ij}){>}\mathrm{E}(Y_{ij}){=}\mu_{ij}$, the assumption of equidispersion of the Poisson model is violated. We assume the overdispersion is introduced by a random subject effect (s_i) which follows a gamma distribution with mean 1 and variance α. The distributions of y_{i0} and y_{i1} conditioned on s_i are then

$$Y_{i0}|s_i \sim \mathrm{Poisson}(s_i \mu_{i0}) \tag{5}$$

$$Y_{i1}|s_i \sim \mathrm{Poisson}(s_i \mu_{i1}), \tag{6}$$

This is the mixed Poisson distribution described by Cook and Wei (2003). Marginalizing on s_i in (6) yields the probability mass function underlying the standard NB model:

$$\mathrm{P}(Y_{i1} = y_{i1}; \mu_{i1}, \alpha) = \frac{\Gamma(y_{i1} + \alpha^{-1})}{\Gamma(y_{i1} + 1)\Gamma(\alpha^{-1})} \left(\frac{1}{1 + \alpha\mu_{i1}}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_{i1}}{1 + \alpha\mu_{i1}}\right)^{y_{i1}}. \tag{7}$$

By incorporating s_i, the variance of Y_{i1} is greater than the mean

$$\mathrm{Var}(Y_{i1}) = \mu_{i1} + \alpha\mu_{i1}^2. \tag{8}$$

In this expression, the term $\alpha\mu_{i1}^2$ provides the extra variance relative to the Poisson model, which is a special case of NB with α approaching 0. NB regression extends the Poisson generalized linear model to account for overdispersion, and incorporates the same, logarithmic link function $\mathrm{g}(\mathrm{E}(Y_{i1})) = \log(\mu_{i1}) = \eta_{i1}$ as Poisson regression, with covariates added to the linear predictor η_{i1}.

Cook and Wei (2003) derived the Conditional Negative Binomial (*CNB*) model from the joint distribution of Y_{i1}, Y_{i0}, and s_i in (5) and (6). Conditioning on the baseline count y_{i0}, the distribution of Y_{i1} follows:

$$\mathrm{P}(Y_{i1} = y_{i1}|y_{i0}; \mu_0, \mu_1, \beta, \alpha) = \frac{\Gamma(y_{i0} + y_{i1} + \alpha^{-1})}{\Gamma(\alpha^{-1} + y_{i0})\Gamma(y_{i1} + 1)} \frac{(1 + \alpha\mu_0)^{\alpha^{-1}+y_{i0}}(\alpha\mu_1)^{y_{i1}}}{(1 + \alpha(\mu_0 + \mu_1))^{\alpha^{-1}+y_{i0}+y_{i1}}}. \quad (9)$$

Note that the *CNB* model can accommodate differing lengths of baseline period (t_{i0}) for each participant, but here it is fixed (t_0).

The estimate of α from the standard NB model is referred to by Hilbe (2011) as the Heterogeneity Parameter (HP). HP reflects the amount of heterogeneity remaining in an NB model. Adding more covariates in η_{i1} may partially explain the heterogeneity from s_i, leading to smaller HP; In comparison, α in the *CNB* model estimates the variance of the underlying gamma mixing distribution described in (5) and (6), and uses information from both y_{i0} and y_{i1} to do so. Larger $\hat{\alpha}$ is indicative of stronger correlation between y_{i1} and y_{i0} due to the subject effect (s_i) they share. Because of the difference in interpretation, we refer to the estimates as HP in NB models following Hilbe, and as $\hat{\alpha}$ in the *CNB* model.

2.2 Including the baseline count as a covariate in a standard NB regression

Although *CNB* is derived from the underlying model for the counts of falls postulated above, it is common practice to include y_{i0} as a covariate in a standard NB model due to its simplicity and accessibility. In this section, we shall investigate how best to set up an NB regression to incorporate the correlation of y_{i0} and y_{i1}.

The conditional expectations of Y_{i0} and Y_{i1} given s_i in (5) and (6) are

$$\mathrm{E}(Y_{i0}|s_i) = \lambda_0 s_i t_0 \quad (10)$$

$$\mathrm{E}(Y_{i1}|s_i, x_i, t_{i1}) = \lambda_1 s_i \exp(\beta x_i) t_{i1}. \quad (11)$$

Hence

$$\mathrm{E}(Y_{i1}|s_i, x_i, t_{i1}) = \frac{\lambda_1}{\lambda_0 t_0} \exp(\beta x_i) \mathrm{E}(Y_{i0}|s_i) t_{i1}. \quad (12)$$

Taking logarithms of both sides yields

$$\log(\mathrm{E}(Y_{i1}|s_i, x_i, t_{i1})) = \log\left(\frac{\lambda_1}{\lambda_0 t_0}\right) + \beta x_i + \log(\mathrm{E}(Y_{i0}|s_i)) + \log(t_{i1}). \quad (13)$$

Substituting y_{i0} for $\mathrm{E}(Y_{i0}|s_i)$ in (13),

$$\log\big(\mathrm{E}(Y_{i1}|y_{i0}, x_i, t_{i1})\big) = \log\left(\frac{\lambda_1}{\lambda_0 t_0}\right) + \beta x_i + \log(y_{i0}) + \log(t_{i1}). \quad (14)$$

As $\log(\lambda_1/(\lambda_0 t_0))$ is a constant, renaming it ζ gives the linear predictor in Poisson/NB regression for Y_{i1} as

$$\log\big(\mathrm{E}(Y_{i1}|y_{i0}, x_i, t_{i1})\big) = g(\mu_{i1}) = \zeta + \beta x_i + \log(y_{i0}) + \log(t_{i1}), \quad (15)$$

suggesting that it may be more appropriate to include the log-transformed y_{i0} as an offset rather than as an untransformed regressor. The combined term $\log(y_{i0}) + \log(t_{i1})$ can be treated as an offset, or if t_{i1} is the same for all subjects, $\log(t_{i1})$ can be incorporated in the constant term and the offset reduced to $\log(y_{i0})$.

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

The performance of the NB model was investigated with the following four linear predictors: (i) ignoring y_{i0}; (ii) including y_{i0} as a covariate; (iii) including $\log(y_{i0})$ as a covariate; (iv) including $\log(y_{i0})$ as an offset. For comparison, results for corresponding Poisson models are also produced. Additionally, Cook and Wei's (2003) *CNB* was included as the benchmark model.

Ignoring the baseline count in the linear predictor (*NB-null*/*Poi-null*): In the NB regression HP accommodates the extra variability in y_{i1} brought about by s_i. Because no explanatory variables (except the intervention indicator and exposure $\log(t_{i1})$) are included, HP estimates α, based on follow-up falls only. The linear predictor is

$$\eta_{i1} = \zeta + \beta x_i + \log(t_{i1}). \tag{16}$$

With this linear predictor, we label the NB model *NB-null* and the Poisson version *Poi-null*. The following models are similarly labelled.

Including the unlogged baseline count in the linear predictor (*NB-unlogged*/*Poi-unlogged*): In *NB-unlogged*/*Poi-unlogged*, the baseline count is included without log-transformation. In *NB-unlogged*, HP still accommodates overdispersion in y_{i1}, but as explained in 2.1, including y_{i0} is likely to result in a smaller estimate of HP than for *NB-null*. The linear predictor for *NB-unlogged*/*Poi-unlogged* is

$$\eta_{i1} = \zeta + \beta x_i + \psi y_{i0} + \log(t_{i1}), \tag{17}$$

where ψ is the coefficient associated with the unlogged baseline count.

Including the logarithm of the baseline count in the linear predictor (*NB-logged*/*Poi-logged*): In accordance with the scaling of y_{i0} in (15), it is included in the linear predictor after log-transformation:

$$\eta_{i1} = \zeta + \beta x_i + \phi \log(y_{i0}) + \log(t_{i1}), \tag{18}$$

with coefficient ϕ. This approach was adopted by Aeberhard *et al.* (2017) for analyzing a falls dataset from a Parkinson's trial. If the logarithmic scale is more appropriate the estimate of HP should be smaller than in *NB-unlogged* as more variability in y_{i1} is likely to be explained. In practice 0.5 is added to all baseline counts to allow transformation when any y_{i0} is zero.

Including the logarithm of the baseline count as an offset in the linear predictor (*NB-offset*/*Poi-offset*): Exactly matching the form of (15), $\log(y_{i0})$ is included as an offset.

$$\eta_{i1} = \zeta + \beta x_i + \log(y_{i0}) + \log(t_{i1}). \tag{19}$$

Again 0.5 is added to all y_{i0} before log transformation. The HP estimate from *NB-offset* is also expected to be smaller than from *NB-unlogged* if log-transformation results in a more appropriate scale for y_{i0}.

2.3 Fitting the models

The models described in sections 2.2 were fitted in `R 3.3.0` using the `negbin` function from the `aod` package for NB models, and the `glm` function for Poisson models. We report P values from Wald tests, estimates and their 95% confidence intervals (CI) and the Akaike Information Criterion (AIC). The `glm.nb` function from the `MASS` package was used to calculate the Cook's distance shown in Figure 2e and 2f. The score test for the NB models was obtained with `st.ml` function from the package `robNB` (Aeberhard, 2016). We reported Anscombe residuals in order to identify outliers following Hilbe's (2011) recommendation. *CNB* models were estimated by the `nlm` function for non-linear minimization, using code made available to us by the authors.

3 Results from Poisson/NB/CNB models fitted to the example dataset

Table 1 shows the results of Poisson, NB, and CNB models fitted to the data from the Goodwin *et al.* (2011) trial. As expected, *Poi-null* results in the largest AIC, while *NB-null* shows an AIC that is an order of magnitude smaller. Although *NB-null* does not incorporate the baseline count, it has lower AIC than any of the Poisson models considered. Compared to *NB-null*, *NB-unlogged* has a smaller AIC (decreased from 931.8 to 844.2) and a smaller HP (decreased from 3.189 to 1.541). The AIC of *NB-logged* further decreases to 744.3, with an FRR (0.698, 95% CI: 0.514 to 0.948) close to that from *NB-unlogged* (FRR=0.677, 95% CI: 0.426 to 1.074). The Wald test of β indicates statistical significance in *NB-logged* (P = 0.021), but not in *NB-unlogged* (P = 0.098). *NB-offset* results in a similar intervention effect (FRR = 0.707, 95% CI: 0.516 to 0.970) to *NB-logged*, and is also statistically significant (P=0.032). $\hat{\phi}$ in *Poi-logged* and *NB-logged* were 1.030 and 0.911 respectively, while $\hat{\psi}$ in *Poi-unlogged* and *NB-unlogged* were 7.02×10^{-3} and 0.019 respectively. *CNB* results in the smallest SE (0.051) of $\hat{\beta}$ and *P* value (< 0.001) amongst the NB based models, and gives an estimate $\hat{\alpha}$ of 2.873.

Diagnostic plots (Figure 2) compare model fit of *NB-unlogged* and *NB-logged*. In Figure 2a, *NB-unlogged* shows a curvilinear pattern between the Anscombe residuals and fitted values. Typically, the model fitted participants with large outcome counts poorly, with predicted much lower than observed counts. After controlling for the log-transformed baseline count, residuals were less skewed, mostly symmetric with respect to zero and followed the underlying distribution (see Figures 2b and 2d). The size of the plotting symbols in Figures 2e and 2f indicate Cook's distance, and while the two participants with the greatest baseline counts were highly influential in the fit of *NB-unlogged*, they do not appear overly influential in *NB-logged*. The larger plotting symbols in Figure 2f reflect inconsistency between low baseline and higher follow-up counts (note - the largest symbols in Figure 2e indicate a greater Cook's distance than symbols of the same size in Figure 2e).

4 Simulation study

4.1 Simulation data sets

Our simulations were based on the findings of the Goodwin *el al.* trial (2011) and falls prevention trials more generally, with some simplifications. 2000 sets of data were simulated in `R 3.3.0` for each scenario considered, using the mixed Poisson distribution described in Section 2.1 (the core code is given in Appendix A.1), with the same number ($n = m/2$) of subjects in the intervention and control groups. Without loss of generality the first n simulated subjects were treated as the control and the second set of n as the intervention group, in which the rate of follow-up falling was adjusted according to parameter β. Following the design of the Goodwin *et al.* trial, we assumed that the length of baseline and follow-up period were the same for all subjects ($t_0 = t_{i1} = t$), so that it was unnecessary to include exposure in the models. The mean baseline count was set at $\mu_{i0} = 30$, close to the observed average baseline count from our motivating dataset of 28. We assumed $\lambda_0 = \lambda_1$ so that $\mu_{i1} = 30$ in the control group. A few sample repeat datasets were examined (not shown) and resembled the pattern observed in Figure 1. The models were fitted in `R` as described in Section 2.3.

We considered scenarios with three levels of intervention effect: β = -0.4 (FRR = 0.67) an intervention effect close to the Goodwin *et al.* data, β = -0.2 (FRR = 0.82) for a smaller effect, and $\beta = 0$ (FRR = 1.00) for checking the empirical type I error rate. The variance of the distribution of subject effects (α) was set at two levels: α = 3 to give a level of overdispersion similar to that in the Goodwin *et al.* data, and $\alpha = 0.5$ to give a lower level of overdispersion. Samples sizes (m) of 50, 100, 200, and 500 were considered. The NB and Poisson models described in Section 2.2 were fitted to each of the 2000 simulated datasets for each scenario, along with the *CNB* model. We recorded $\hat{\beta}$ and their model-based standard error (SE), $\mathrm{SE}(\hat{\beta})$, from the model fits to each simulated dataset. The following statistics (White, 2010) were calculated:

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

$$\widehat{\text{Bias}} = \text{av}(\hat{\beta}) - \beta, \tag{20}$$

where $\text{av}(\hat{\beta})$ is the average (av) of the estimates of β across repeats for each scenario; and the SE of $\widehat{\text{Bias}}$ obtained from the simulations, the "Monte Carlo error" (MCError), defined as:

$$\text{MCError}(\widehat{\text{Bias}}) = \frac{\text{EmpSE}}{\sqrt{\text{n}_{\text{sim}}}}, \tag{21}$$

where n_{sim} is the number estimates obtained for each scenario, and the empirical SE (EmpSE) of $\hat{\beta}$ is calculated as the standard deviation of the $\hat{\beta}$. The ModSE is defined as the average of the model-based SE, $\text{SE}(\hat{\beta})$. The relative error is:

$$\text{Relative Error} = \frac{\text{ModSE}}{\text{EmpSE}} - 1, \tag{22}$$

The following average estimates across the repeats for each scenario were also recorded: $\text{av}(\text{HP})$, $\text{av}(\hat{\alpha})$, $\text{av}(\hat{\psi})$, and $\text{av}(\hat{\phi})$. Repeats where the algorithm failed to converge were excluded. We also excluded estimates that converged to an incorrect value judged by $|\hat{\beta} - \beta| > 5$ or $\text{SE}(\hat{\beta}) > 1$ (the selection criteria were chosen by examining the distribution of $\hat{\beta}$ and model-based SE). The proportion of significant results from the Wald test of β amongst the replicates is the empirical power when $\beta \neq 0$ and type I error rate otherwise. We also examined the empirical power and type I error rates of score tests for the NB models.

To examine the sensitivity of results to the addition 0.5 to y_{i0} in *NB-logged* and *NB-offset*, we conducted simulations to investigate alternative values (0.01, 0.1, and 1).

4.2 Simulation results

Algorithms converged to appropriate solutions in most cases: the numbers of replicates included for each scenario are shown in Table S1. Figure 3 compares the NB/CNB models in terms of $\widehat{\text{Bias}}$ with the 95% confidence intervals calculated from the MCError. In most scenarios, $\hat{\beta}$ in *NB-null* are close to the underlying value. The relative error of $\hat{\beta}$ is within ±6% (Figure 4), that the model-based SE is a good estimator of the variance of $\hat{\beta}$. The average HP from *NB-null* is not far from the underlying α (Table S1). *NB-null* has the lowest power (Figure 5a) in all scenarios amongst all the NB and CNB models that we considered, although its type I error rates (Figure 5b) are close to the nominal level (0.05). *NB-unlogged* generally has improved power compared to *NB-null* due to the extra information contributed by y_{i0}, but this improvement is not as great when $\alpha = 3$ as when $\alpha = 0.5$. The type I error rates of *NB-unlogged* are lower than the assigned level 0.05, especially when $\alpha = 3$ (around 0.015), and the pattern of low power is consistently seen even for large sample sizes when β = -0.2.

By log-transforming the baseline count, *NB-logged* achieves greater power (Figure 5a) than *NB-unlogged* in all scenarios. Although slightly inflated type I error rates (Figure 5b) are shown in smaller sized scenarios (maximum being 0.071 when the sample size is 50 and $\alpha = 3$), the rates converge to the nominal level as the sample size increases. Results from *NB-offset* are close to *NB-logged* with almost identical power, but the type I error rates of *NB-offset* are marginally closer to 0.05 when α = 3. As shown in Figure S1, the type I error rates of the score test in both *NB-logged* and *NB-offset* are close to the nominal level when the sample size is small ($m = 50$). For *NB-unlogged*, when $\alpha = 3$ the type I error rates are too high at smaller m but too low at larger m.

In *NB-unlogged*, $\widehat{bias}$ has smaller MCErrors compared to *NB-null* (Figure 3), suggesting smaller variation in $\hat{\beta}$, but the model-based SEs are typically overestimated (Figure 4). In comparison, the model-based SEs in *NB-logged* and *NB-offset* are close to the empirical values. These models also

result in smaller HP (Table S1). The average of the $\hat{\phi}$s in *NB-logged* are close to 1, generally being greater than 1 when β = 3 and smaller than 1 when β = 0.5 (Table S1). It was found that adding different values to y_{i0} before log transformation had little impact on the estimation of β (Table S2), or the power and type I error rates of the related Wald test (Table S3).

The power of *CNB* (Figure 5a) is equal to or higher than all the NB models we considered, although it only slightly outperforms *NB-logged* and *NB-offset*. When β = -0.2, *NB-logged*, *NB-offset*, and *CNB* show larger improvements in terms of power compared to *NB-unlogged* and *NB-null* than the scenarios with β = -0.4. The type I error rates of *CNB* (Figure 5b) are stable and close to 0.05 in all scenarios. Model-based SEs from *CNB* are close to their corresponding empirical value (Figure 4), and on average, the estimates, $\hat{\alpha}$, are close to the underlying α (Table S1).

The model-based SEs of $\hat{\beta}$ in Poisson models (Figure S3) are substantially underestimated and there are excessively high type I error rates (Figure S4b), especially in scenarios with larger α (3).

5 Discussion

The NB regression model is widely used in the analysis of falls prevention trials. When a baseline count of falls is available, the question remains how best to utilize it in modeling? One approach is to model the distribution conditioned on the baseline count using Cook and Wei's (2003) *CNB* model. We simulated data from the mixed Poisson distribution underlying their model, and *CNB* generally performed the best among all the considered models as was anticipated, with the highest power and type I error rates close to 0.05 even for the smallest sample sizes. However, it is not available in most statistical packages and thus difficult to use in practice. NB regression, in comparison, is increasingly popular and easily fitted in dedicated commands as a mixed Poisson models. The question remains whether an NB model can approximate *CNB* without loss of power.

NB-null, ignoring the baseline count, resulted in lower power than the NB models including the count in any way. Although the NB model including the untransformed baseline count (*NB-unlogged*) does not reflect the scaling of the model, it outpowered *NB-null* in all scenarios. The variance of the underlying mixing distribution, parameter α, controls the potential for the follow-up count to be explained by the baseline count. Varying α was found to impact on the disparity in power between *NB-null* and *NB-unlogged*: the power gain of *NB-unlogged* over *NB-null* was not as great for the larger α (3) compared to the smaller α (0.5) we considered. A further problem was indicated: the type I error rates of *NB-unlogged* were noticeably deflated when $\alpha = 3$ and did not appear to converge to 0.05 as the sample size increased.

The NB models including the log-transformed baseline count as a covariate (*NB-logged*) or as an offset (*NB-offset*) had lower variability in $\hat{\beta}$ than in NB-unlogged, and the empirical SEs were more closely approximated by model-based SEs. Both *NB-logged* and *NB-offset* yielded great improvement over *NB-unlogged* with substantially increased power in the Wald tests of β, and small, almost identical levels of bias, along with similar HPs. They were both only slightly less powerful than *CNB*, and any disparity diminished with increasing sample size. In *NB-logged*, the average estimated coefficients for the logged baseline count were generally close to 1. We may conclude that the approach of fitting the log-transformed baseline count as an explanatory variable, better reflects the scaling of its relationship to the underlying falls rate.

In broad terms Hilbe's (2011) Heterogeneity Parameter (HP) from an NB regression fit reflects how much variability remains unexplained by the linear predictor. HPs from *NB-null* were close to the underlying α, meaning that heterogeneity was accommodated in the model, but not explained by any of the covariates. The HP was greatly reduced after accounting for the baseline count in *NB-unlogged*, and it was further reduced in *NB-logged* and *NB-offset*, reflecting the greater explanatory power of the log-transformed compared to untransformed baseline count. Poisson models do not incorporate variability over participants and are known to result in underestimation of model-based SE and inflated

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

type I error rate for β (Fitzmaurice, 1997). One might argue that because inter-subject variability is shared in the baseline and follow-up counts, by including the logged baseline count in a Poisson regression, overdispersion would be eliminated so that NB regression would be unnecessary. Our simulation showed this not to be the case: (i) type I error rates from *Poi-logged* and *Poi-offset* were too high at around 0.16; and (ii) the AIC from *Poi-logged* and *Poi-offset* were larger than that from the most basic NB model (*NB-null*) fitted to our motivating dataset.

The type I error rates of *NB-logged* and *NB-offset* were moderately inflated at smaller sizes. The rates converged to nominal level as the sample size increased. Aban *et al.* (2009) reported inflated type I error rates for Wald tests in NB regression when sample sizes were small (under size 200 in total). Aeberhard *et al.* (2017) also discussed this issue, and recommended using the robust TETT (Tilted Exponential Tilting Test) for NB based hypothesis tests with small sample sizes. To our knowledge, this test can only be obtained from the author's `R` package `robNB` (Aeberhard, 2016). We focused on evaluating the performance of Wald tests: being the default in widely available NB regression commands, it is most commonly used in practice. With sample sizes of 200 the type I error rates were only slightly higher than target. Our simulation study was carried out for trials of size 50, 100, 200 and 500 (totaled across both arms). Though performance would have improved further with larger trial sizes, the range we considered covers the majority (80%) of falls prevention trials included in the 2012 Cochrane review (Gillespie *et al.*, 2012). Our simulations of the score test suggested that it is not overly liberal when sample size is small, but it is not available in most commands for NB modeling.

Our motivating dataset bore out our conclusions from the simulations: *NB-unlogged* achieved substantially smaller AIC than *NB-null* by including the untransformed baseline count. *NB-logged* and *NB-offset* further decreased AIC, and resulted in significant results (P=0.021 and 0.032 respectively) compared to *NB-unlogged* (P=0.098). When data were greatly overdispersed, *NB-unlogged* failed to cope effectively with large follow-up counts: this was born out by the large Cook's distances for participants with large baseline and follow-up counts when *NB-unlogged* was fitted, while this issue did not appear for *NB-logged*. The significant Wald test result found in the NB-logged model may reflect the liberal nature of the test at relatively small sample sizes, but CNB also yielded a significant test result and does not suffer from the same problem with small sample sizes.

All the models discussed here can accommodate differing lengths of follow-up, but how to accommodate differing lengths of baseline in NB regression remains an issue. The baseline count over a short period will be a poorer measure of a participant's underlying rate of falling, and the shorter the length of the baseline, the more likely the participant is to report a zero count. We adopted the pragmatic approach of adding the value of 0.5 before log-transformation so that participants reporting zero baseline counts were included. The choice of value to add is a trade-off: adding a smaller value results in less change on the untransformed scale, but after log-transformation leads to large negative values. Our simulations show that the values we considered (+0.01, +0.1, +0.5, and +1) do not substantially affect the estimation of β. We chose 0.5, in keeping with standard continuity corrections. Cook and Wei's (2003) conditional model, *CNB*, uses information in the baseline count to improve the estimate of α, taking account of the precision available from each participant's length of baseline, and no difficulty arises with zero counts.

Our study demonstrated that very little is lost by using the standard NB model with the baseline count included as a log-transformed regressor, compared to the *CNB* model. Including the baseline count of falls greatly increases power, thus a baseline count should be collected in falls prevention trials, as is generally recognised (Assmann *et al.*, 2000). The NB model including the log-transformed baseline count as a regressor or offset may be fitted in `R`, `SAS`, `Stata`, `SPSS` and probably other statistical packages. With moderate to larger sample sizes, the power of *NB-logged* and *NB-offset* model are only slightly less than *CNB*, and the type I error rate are both close to the nominal significant level of 0.05. This approach is easy to implement and widely accessible to medical researchers.

Acknowledgements VG is funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. We thank a referee for helpful comments on an earlier draft.

Conflict of Interest

The authors have declared no conflict of interest

Biometrical Journal **52** (2010) 61, ZZZZZZ / **DOI** 10.1002/bimj.201010000

Appendix

A.1 The R code for simulation

The R code of simulating baseline and follow-up counts from mixed Poisson distribution

```
obs <- 10000
alpha <- 3

## Simulate the gamma-shaped subject effect
s <- rgamma(obs, shape = 1/alpha, scale = alpha)

## Simulate the baseline counts
mu <- 20
lambda_baseline <- mu * s
y_0 <- rpois(n = obs, lambda = lambda_baseline)

## Simulate the group allocation
group <- c(rep(1, obs/2), rep(0, obs/2))
beta <- -0.2

## Simulate the follow-up counts
lambda_followup <- exp(beta * group) * mu * s
y_1 <- rpois(n = obs, lambda = lambda_followup)
```


References

Aban, I. B., Cutter, G. R. and Mavinga, N. (2009). Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. *Computational Statistics & Data Analysis*, 53(3), pp. 820–833. doi: 10.1016/j.csda.2008.07.034.

Aeberhard, W. H., Cantoni, E. and Heritier, S. (2017). Saddlepoint tests for accurate and robust inference on overdispersed count data. *Computational Statistics & Data Analysis*, 107(October), pp. 162–175. doi: 10.1016/j.csda.2016.10.009.

Aeberhard W. H. (2016). robNB: Robust estimation and tests for negative binomial regression. R package version 0.2. https://github.com/williamaeberhard/robnb

Assmann, S. F., Pocock, S. J., Enos, L. E. and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209), pp. 1064–1069. doi: 10.1016/S0140-6736(00)02039-0Cite.

Cook, R. J. and Wei, W. (2003). Conditional analysis of mixed Poisson processes with baseline counts: implications for trial design and analysis. *Biostatistics (Oxford, England)*, 4(3), pp. 479–494. doi: 10.1093/biostatistics/4.3.479.

Donaldson, M. G., Sobolev, B., Cook, W. L., Janssen, P. a. and Khan, K. M. (2009). Analysis of recurrent events: A systematic review of randomised controlled trials of interventions to prevent falls. *Age and Ageing*, 38(December 2008), pp. 151–155. doi: 10.1093/ageing/afn279.

Fitzmaurice, G. M. (1997). Model selection with overdispersed data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1), pp. 81–91. doi: 10.1111/1467-9884.00061.

Gillespie, L. D., Robertson, M. C., Gillespie, W. J., Sherrington, C., Gates, S., Clemson, L. M. and Lamb, S. E. (2012). Interventions for preventing falls in older people living in the community, in Gillespie, L. D. (ed.) *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd, p. CD007146. doi: 10.1002/14651858.CD007146.pub3.

Glynn, R. J. and Buring, J. E. (1996). Ways of measuring rates of recurrent events. *BMJ (Clinical research ed.)*, 312(February), pp. 364–367. doi: 10.1136/bmj.312.7027.364.

Goodwin, V. A., Richards, S. H., Henley, W., Ewings, P., Taylor, A. H. and Campbell, J. L. (2011). An exercise intervention to prevent falls in people with Parkinson's disease: a pragmatic randomised controlled trial. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(11), pp. 1232–1238. doi: 10.1136/jnnp-2011-300919.

Hauer, K., Lamb, S. E., Jorstad, E. C., Todd, C. and Becker, C. (2006). Systematic review of definitions and methods of measuring falls in randomised controlled fall prevention trials. *Age and Ageing*, 35(1), pp. 5–10. doi: 10.1093/ageing/afi218.

Hilbe, J. M. (2011). *Negative Binomial Regression*. 2nd edn, *Cambridge University Press*. 2nd edn. doi: 10.1111/j.1540-6210.2010.02207.x.

White, I. R. (2010). *simsum: Analyses of simulation studies including Monte Carlo error*, Stata Journal, 10(3), p. 369.

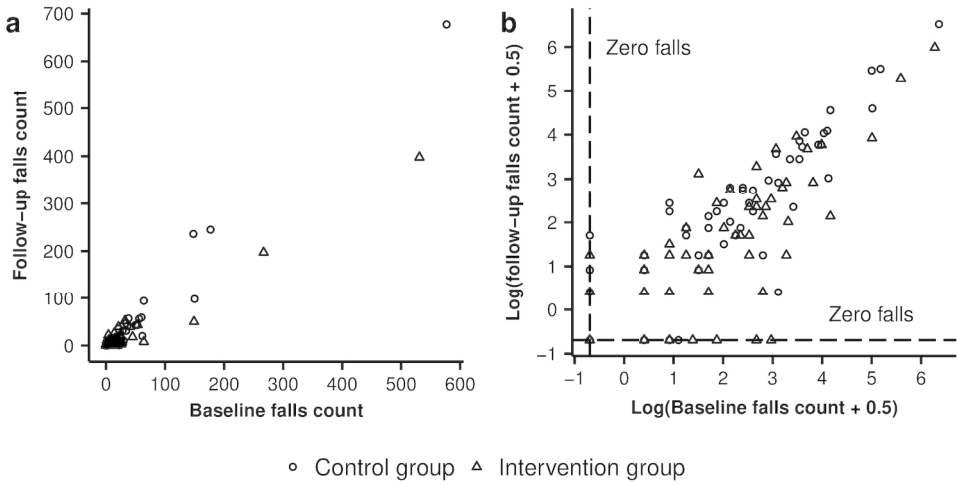

Figure 1 Follow-up against baseline counts of falls reported by Goodwin et al. (2011) (n=124, Spearman =0.813, P<0.001). 0.5 is added to all counts before log-transformation in order to include zero counts in 1b.

150x75mm (300 x 300 DPI)

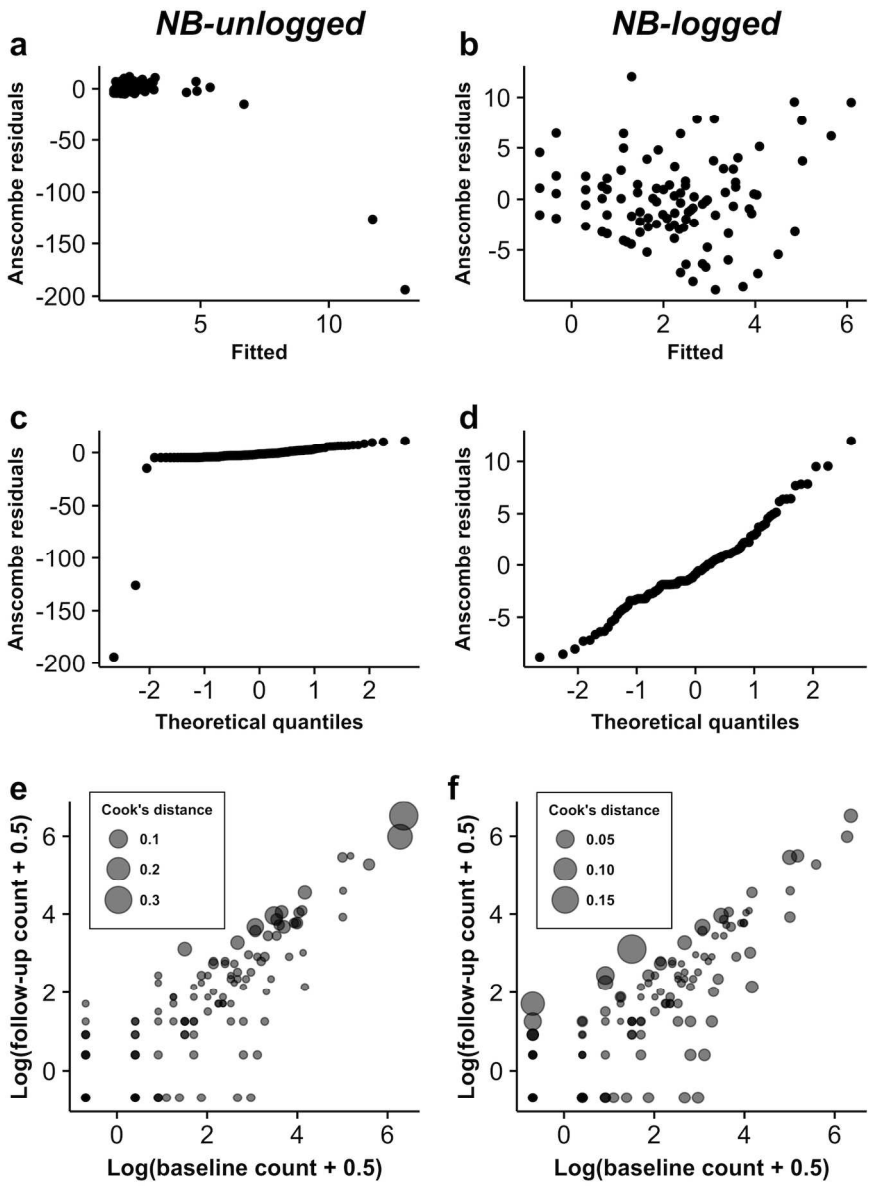

Figure 2 Diagnostic plots from NB-unlogged (Column 1) versus NB-logged (Column 2) fitted to the Goodwin et al. (2011) data. (a-b) Anscombe residuals versus fitted. (c-d) Normal Q-Q plot of Anscombe residuals. (e-f) Follow-up versus baseline count in logarithmic scale (0.5 is added before the log-transformation in order to include zero counts), with the size of plotting symbols indicating Cook's distance.

150x199mm (300 x 300 DPI)

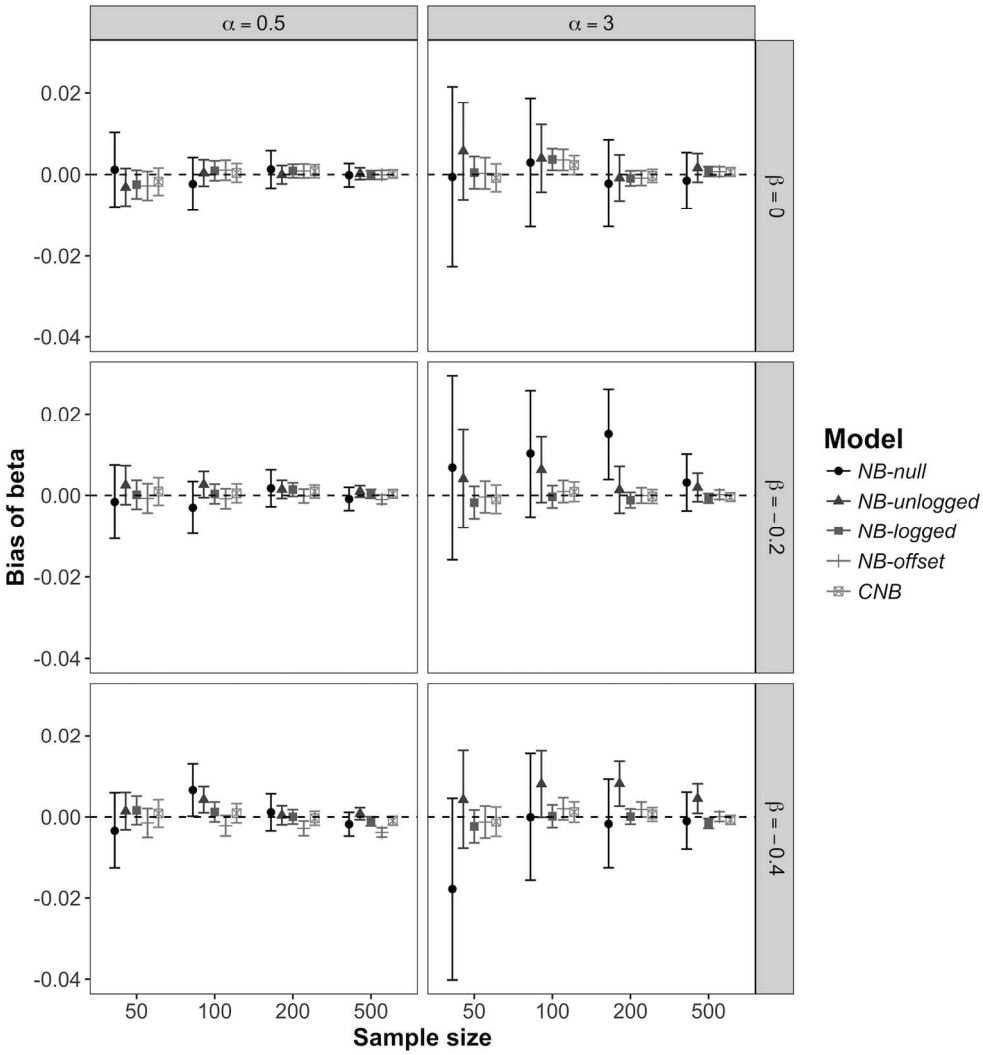

Figure 3 Bias plot of NB and CNB models. Each point indicates the bias of the estimation of β (the average of estimates minus the true value), with error bars showing the 95% confidence interval calculated using the MCError.

150x159mm (300 x 300 DPI)

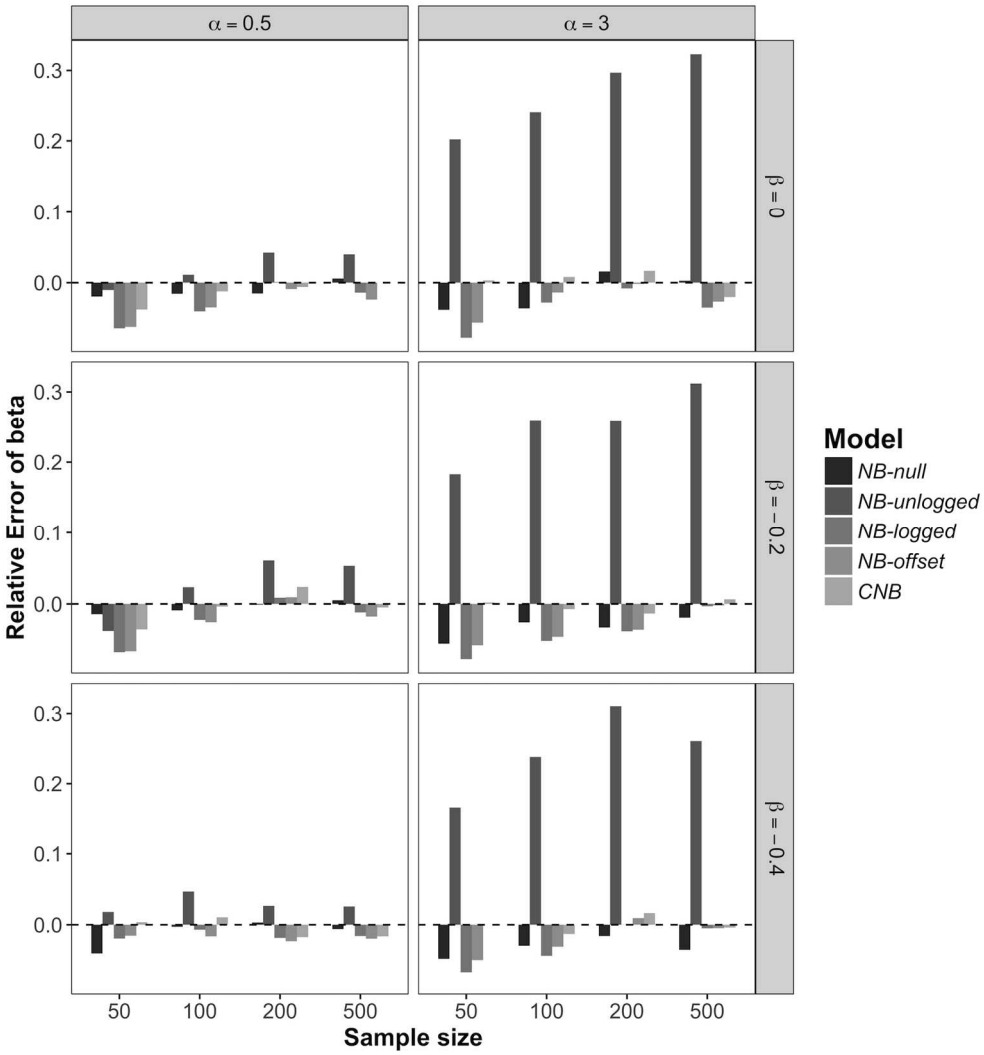

Figure 4 Relative error plot of NB and CNB models.

150x159mm (300 x 300 DPI)

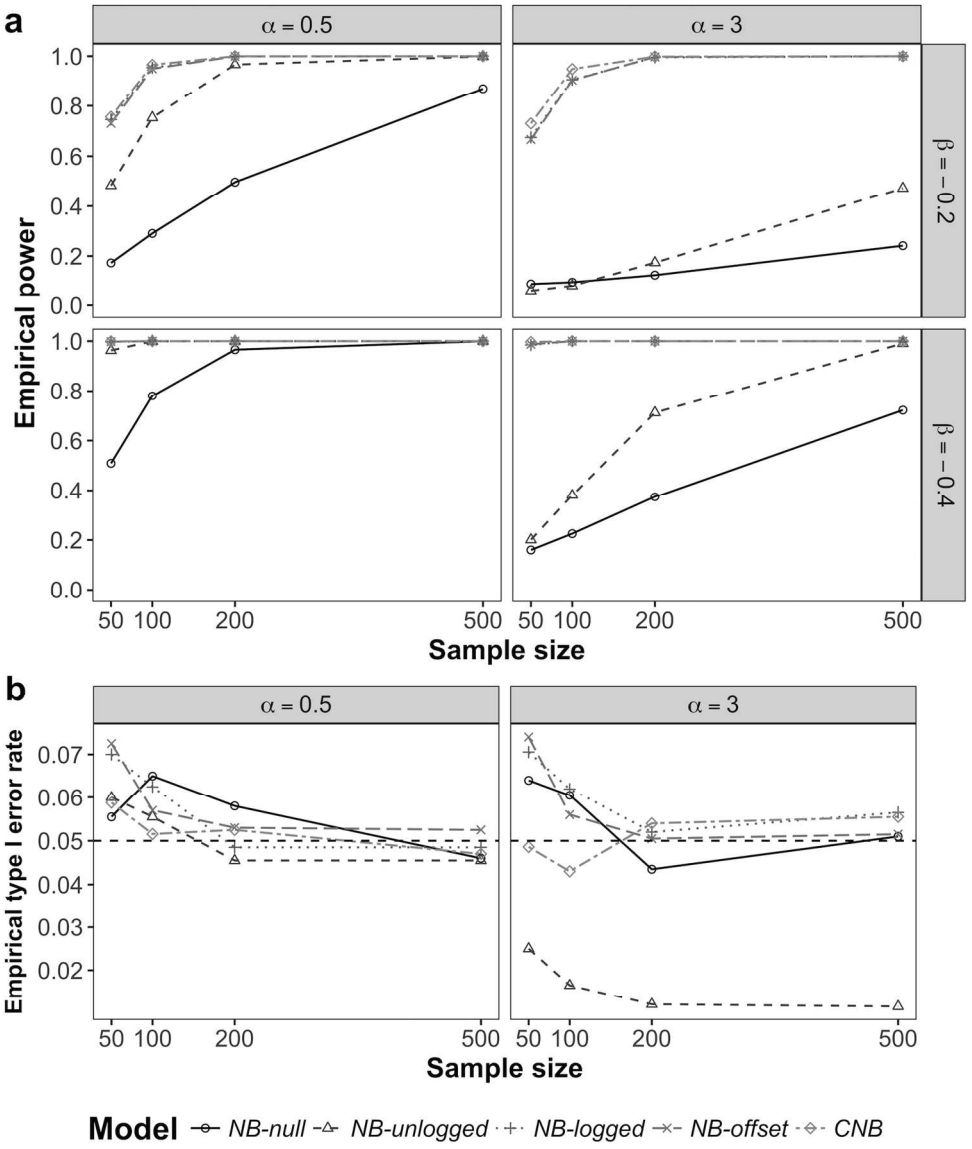

Figure 5 Simulation results of the Wald test from NB and CNB models. (a) Empirical Power; (b) Empirical type I error rates.

150x182mm (300 x 300 DPI)

Table 1 Summary of Poisson/NB/CNB models fitted to the Goodwin *et al.* data (n=124).

Model	AIC	$\widehat{\boldsymbol{\beta}}$ (SE)	FRR (95%)	P	$\widehat{\boldsymbol{\psi}}$ (SE)	$\widehat{\boldsymbol{\phi}}$ (SE)	HP
Poi-null	9996.1	-0.571 (0.037)	0.565 (0.525, 0.608)	< 0.001			
Poi-unlogged	3247.6	-0.472 (0.038)	0.624 (0.580, 0.672)	< 0.001	7.02×10^{-3} (6.78×10^{-5})		
Poi-logged	1131.5	-0.480 (0.037)	0.619 (0.575, 0.666)	< 0.001		1.030 (0.012)	
Poi-offset	1135.6	-0.479 (0.037)	0.619 (0.577, 0.666)	< 0.001			
NB-null	931.8	-0.572 (0.323)	0.565 (0.300, 1.064)	0.077			3.189
NB-unlogged	844.2	-0.391 (0.236)	0.677 (0.426, 1.074)	0.098	0.019 (0.004)		1.541
NB-logged	744.3	-0.359 (0.156)	0.698 (0.514, 0.948)	0.021		0.911 (0.048)	0.511
NB-offset	745.5	-0.346 (0.161)	0.707 (0.516, 0.970)	0.032			0.519
							$\hat{\alpha}$
CNB		-0.479 (0.051)	0.619 (0.561, 0.684)	< 0.001			2.873