A Simple Approach to Sample Size Calculation for Count Data in Matched Cohort Studies

Dexiang Gao[a,b,*], Gary K. Grunwald[b] and Stanley Xu[b,c]

[a]*Department of Pediatrics, School of Medicine, University of Colorado Denver, USA*

[b]*Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Colorado, USA*

[c]*Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA*

Abstract: In matched cohort studies exposed and unexposed individuals are matched on certain characteristics to form clusters to reduce potential confounding effects. Data in these studies are clustered and thus dependent due to matching. When the outcome is a Poisson count, specialized methods have been proposed for sample size estimation. However, in practice the variance of the counts often exceeds the mean (i.e. counts are overdispersed), so that Poisson methods don't apply. We propose a simple approach for calculating statistical power and sample size for clustered Poisson data when the proportion of exposed subjects in a cluster is constant across clusters. We extend the approach to clustered count data with overdispersion, which is common in practice. We evaluate these approaches with simulation studies and apply them to a matched cohort study examining the association of parental depression with health care utilization. Simulation results show that the methods for estimating power and sample size performed reasonably well under the scenarios examined and were robust in the presence of mixed exposure proportions up to 30%.

Keywords: Clustered Poisson data, Overdispersion, Subject heterogeneity, Statistical power, Sample size.

1. INTRODUCTION

Count data appear frequently in randomized clinical trials and in observational studies. Interest is often in comparing incidence rates among two or more groups. For example, in a retrospective study, Kornek *et al*. [1] compared annual relapse rates, counts of new T2/fluid-attenuated inversion recovery lesions and contrast-enhancing lesions on magnetic resonance imaging, and counts of adverse events, between patients with and without natalizumab treatment. Rothman and Greenland [2] and Graham *et al*. [3] compared breast cancer incidence rates for subjects exposed to X-ray fluoroscopy during treatment for tuberculosis to those not exposed. In some studies, subjects may form clusters. A common example is a matched cohort design, where exposed subjects are matched to unexposed subjects to form clusters. The matching is formed either by design in randomized clinical trials such as a hospital or a health care provider defining a cluster, where exposure is randomized within clusters, or in observational studies by matching exposed and unexposed individuals on covariates to reduce confounding effects [4,5].

Several quantities are involved in matched cohort studies, including the number of clusters, the number of subjects per cluster (cluster size), and the proportion of exposed subjects in a cluster (exposure proportion). Studies with constant or approximately constant exposure proportion across clusters are common. An example of clustered count data is the study by Sills *et al*. [6], who used a matched cohort design to examine the association between parental depression and children's health care use. The exposed children were 0-17 years of age, enrolled in Kaiser Permanente of Colorado for at least six months during a study period of July 1997 to December 2002. They were linked with at least one parent/subscriber with depression diagnosis (exposure). Unexposed children were selected from a pool of similar children whose parents did not have a depression diagnosis and were matched to exposed children on age, gender, membership eligibility, and enrollment period. Among exposed children, 85.6% had two matched controls, the remaining had one matched control. Outcome measures included number of clinic visits of any type during the enrollment period, which was obtained from the child's payment files and electronic medical charts.

Sample size and power estimation are not straightforward in planning matched cohort studies with count outcome due to dependence of subjects within a cluster and possible overdispersion of count outcome. Several authors considered sample size estimation for studies with Poisson outcomes where the mean and variance of count outcome are equal. Among them, Ng and Tang [7] examined four methods for comparing

*Address correspondence to this author at the Department of Pediatrics, SOM, Department of Biostatistics and Informatics, CSPH, University of Colorado Denver, 12477 E. 19th Avenue, Aurora, CO 80045, USA; Tel: 303-724-4356; Fax: 303-724-4489; E-mail: dexiang.gao@ucdenver.edu

E-ISSN: 1929-6029/14

means of two independent Poisson samples. Amatya [8] developed methods for sample size estimation in matched cohort studies with Poisson outcomes, and demonstrated that their methods performed better than similar methods previously proposed. These previous results provide the necessary methods in studies where the count outcome follows a Poisson distribution conditional on random cluster effects. However, in practice count data tend to be overdispersed (i.e. the variance is greater than the mean) [9-12]. Friede and Schmidli [13] extended the results of Ng and Tang and developed sample size and power estimation for independent count data with overdispersion.

We are not aware of methods for estimating sample size and power for exposure effects in matched cohort studies where count outcomes are also overdispersed. In this paper, we develop and evaluate a simple approach for sample size and power estimation for an exposure effect in matched cohorts studies with a count outcome overdispersed under the assumption of constant exposure proportion. We consider two scenarios based on whether the conditional distribution of the count outcome given random cluster effects and covariates is just Poisson or overdispersed. We refer to these data as clustered Poisson data and clustered count data with overdispersion, respectively.

In Sections 2.1, we started with Ng and Tang's analytic methods for comparing means of two independent Poisson samples [7]. We then provide a general expression for power and sample size for count data in Section 2.2 and emphasize that the key is to obtain the variance of $\hat{\beta}_1$, ($\mathrm{Var}(\hat{\beta}_1)$), where $\hat{\beta}_1$ is maximum likelihood estimate (MLE) of the coefficient of an exposure effect, referred to as the log rate ratio. In Section 2.3 we provide an asymptotic formula for $\mathrm{Var}(\hat{\beta}_1)$ for clustered Poisson data with a constant exposure proportion across all clusters. In Section 2.4, we first discuss the $\mathrm{Var}(\hat{\beta}_1)$ for independent count data with overdispersion, and then propose an asymptotic formula for $\mathrm{Var}(\hat{\beta}_1)$ for clustered count data with overdispersion. In Section 3 we use simulations to evaluate the performance of these formulas across a range of parameter values. We also assess the robustness of our approach by allowing a portion of clusters to have different exposure proportions. Section 4 illustrates application of these methods to the Kaiser study of association between parental depression and children's health care use. We close the paper with some discussions in Section 5.

2. STATISTICAL METHODS

2.1. Comparing Means of Two Independent Poisson Samples

For independent Poisson data Y_j, assume $Y_j \big| x_j$ ~ Poisson (μ_j) with $\mu_j = e^{\beta_0+\beta_1 x_j}$ where x_j is the indicator for exposure (1 for exposed and 0 for unexposed) for subject $j, j = 1,\ldots,M$, β_0 is the coefficient determining the log incidence rate for the unexposed group, and β_1 is the coefficient of the exposure effect. Ng and Tang [7] compared four test statistics for testing equality of two Poisson means, $\mu_0 = \mu_1$ (i.e., β_1=0) where $\mu_0 = e^{\beta_0}$ and $\mu_1 = e^{\beta_0+\beta_1}$. Among the four test statistics, the Wald test statistic,

$$W = \frac{\hat{\beta}_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} \tag{1}$$

was found to have a robust type I error rate (0.04-0.06) and to have empirical power close to the pre-chosen power level based on simulation studies. Let n_0 and n_1 be the numbers of unexposed and exposed subjects with $n_0 + n_1 = M$, then $\mathrm{Var}(\hat{\beta}_1)$ for independent Poisson data, obtained using a delta method, is

$$\mathrm{Var}(\hat{\beta}_1) = V_P(\hat{\beta}_1) = \frac{1}{n_0\mu_0} + \frac{1}{n_1\mu_1} \tag{2}$$

where subscript P in $V_P(\hat{\beta}_1)$ represents Poisson data.

2.2. A General Expression for Power and Sample Size

Because of the established theoretical and empirical properties of the Wald statistic and its good performance for testing equality of means for independent Poisson samples [7], we base our methods on the Wald statistic and use the following general statistical power formula for testing β_1 = 0 versus β_1 = β_1^* for count data:

$$\mathrm{Power} = \Phi\left(Z_{\alpha/2} + \frac{\beta_1^*}{\mathrm{Var}(\hat{\beta}_1)} \right) \tag{3}$$

where Φ is the standard normal cumulative distribution function (CDF), $Z_{\alpha/2}$ is the $\alpha/2$ quantile value of the standard normal distribution, and β_1^* is the hypothesized alternative (i.e. detectable effect) value. Given the sample sizes n_0 and n_1 (or the total sample size and the exposure proportion), the unexposed group incidence rate (μ_0), and the log rate ratio to be detected (β_1^*), statistical power can be calculated from (3). Given β_0, β_1^*, statistical power, and exposure proportion, n_0 and n_1 can be calculated by inverting (3).

The form of $\mathrm{Var}(\hat{\beta}_1)$ used in (3) depends on whether count data are independent or clustered, and whether they are conditionally Poisson distributed or overdispersed. For independent Poisson data this variance is given in equation (2); for clustered and/or overdispersed count data, formulas for $\mathrm{Var}(\hat{\beta}_1)$ are given below.

2.3. Asymptotic Variance of $\hat{\beta}_1$ for Clustered Poisson Data

Let Y_{ij} denote the count outcome for the jth subject in the ith cluster where j =1,…, n, i=1,…, N, and the exposure indicator variable x_{ij} is 1 for exposed and 0 for unexposed subjects. In this paper we employed a statistical model that assumes a common random cluster effect for all subjects within a cluster to analyze clustered count data [14-16]. With a log link function, for clustered Poisson data we have

$$Y_{ij}|\gamma_i \sim \text{Poisson}\,(\lambda_{ij}), \lambda_{ij} = e^{\beta_0+\beta_1 x_{ij}+\gamma_i} \tag{4}$$

where γ_i is a cluster-specific random effect independent across clusters. The inclusion of γ_i induces a correlation among the outcome measures within a cluster. If γ_i is assumed to be $N(0,\sigma^2)$, then marginal (unconditional on γ_i) moments are $E(Y_{ij}) = \mu_{ij} = e^{\beta_0+\beta_1 x_{ij}+\sigma^2/2}$ and $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}^{\ 2}(e^{\sigma^2}-1)$. The second term $\mu_{ij}^{\ 2}(e^{\sigma^2}-1)$ in $\mathrm{Var}(Y_{ij})$ is due to clustering. Note that we use λ for means conditional on random cluster effects and μ for marginal means. It has been shown that for clustered Poisson (CP) data with a constant exposure proportion for all clusters, the asymptotic variance of $\hat{\beta}_1$ can be obtained in a closed form [14,15]:

$$\mathrm{Var}(\hat{\beta}_1) = V_{CP}(\hat{\beta}_1) = \frac{1}{Nn}\left[\frac{1}{(1-R)\mu_0} + \frac{1}{R\mu_1}\right] \tag{5}$$

where N is number of clusters and n is cluster size, R is the exposure proportion, $\mu_0 = e^{\beta_0+\sigma^2/2}$ and $\mu_1 = e^{\beta_0+\beta_1+\sigma^2/2}$. Subscript CP in $V_{CP}(\hat{\beta}_1)$ represents clustered Poisson data.

2.4. Asymptotic Variance of $\hat{\beta}_1$ for Clustered Count Data with Overdispersion

2.4.1. Independent Overdispersed Count Data

Count data are often overdispersed due to unobserved subject heterogeneity [11,17-19]. A number of statistical models have been proposed to accommodate overdispersion [20-26]. Overdispersed count data can be modeled by assuming $\mathrm{Var}(Y) = \phi\mu$ where $\phi > 1$ is called the scale or dispersion parameter. Friede and Schmidli [13] developed sample size and power estimation for independent count data with overdispersion by replacing the variance in (1) with $V_0(\hat{\beta}_1) = \phi V_0(\hat{\beta}_1)$, where subscript O in $V_0(\hat{\beta}_1)$ represents overdispersed count data. With simulation they demonstrated that their approach gave good approximations to target power.

Another approach for analyzing count data with overdispersion is to allow random variation in the conditional mean by introducing a subject-specific random multiplicative term θ_j; overdispersion parameter ϕ depends on the distribution of θ_j. Note that subscript j is for subject, and subscript i for cluster is not needed here for independent count data. It is assumed in general that θ_j is independent across subjects and has a known parametric distribution with mean 1. A common choice is θ_j ~ gamma $(1/\tau\,,\tau\,)$ so that $\mathrm{E}(\theta_j) = 1$ and $\mathrm{Var}(\theta_j) = \tau$. Then

$$Y_j|\theta_j \sim \text{Poisson}\,(\pi_j),\ \pi_j = \theta_j e^{\beta_0+\beta_1 x_j} \tag{6}$$

and

$$\mathrm{Var}(Y_j) = \mu_j + \tau\mu_j^2 = \mu_j(1+\mu_j\tau) \tag{7}$$

where μ_j = exp($\beta_0 + \beta_1 x_j$) [24], the overdispersion parameter $\phi = (1+\mu_j\tau)$. In this paper, we use the model in (6) and the variance form in (7) to extend the approach of Friede and Schmidli [13] to clustered count data with overdispersion as detailed in the following section. For model (6), the variance of $\hat{\beta}_1$ for independent overdispersed count data is

$$\mathrm{Var}(\hat{\beta}_1) = V_0(\hat{\beta}_1) = \frac{1}{Nn}\left[\frac{\phi_0}{(1-R)\mu_0} + \frac{\phi_1}{R\mu_1}\right]$$

where $\phi_0 = 1 + \tau e^{\beta_0}$ for unexposed and $\phi_1 = 1 + \tau e^{\beta_0+\beta_1}$ for exposed.

2.4.2. Clustered Count Data with Overdispersion

When both overdispersion and clustering are present in count data, models (4) and (6) can be extended to

$$Y_{ij}\left|(\theta_{ij},\gamma_i) \sim \text{Poisson}\,(\lambda_{ij}),\ \lambda_{ij}=\theta_{ij}e^{\beta_0+\beta_1 x_{ij}+\gamma_i}\right. \qquad (8)$$

where θ_{ij}, assumed to have $E(\theta_{ij})=1$, represents overdispersion due to subject heterogeneity, and γ_i represents random cluster effects [27]. There is no closed form solution for $\mathrm{Var}(\hat{\beta}_1)$ for model (8). However, it has been shown [27] that the variance of Y_{ij} is

$$\mathrm{Var}(Y_{ij})=\mu_{ij}+\mu_{ij}^2(e^{\sigma^2}-1)+\mu_{ij}^2\tau e^{\sigma^2} \qquad (9)$$

where $\mu_{ij}^2(e^{\sigma^2}-1)$ is the extra variance due only to clustering but not overdispersion (see Section 2.3 for clustered Poisson data). We propose to modify equation (5) by applying overdispersion parameters to obtain the variance of $\hat{\beta}_1, V_{CO}(\hat{\beta}_1)$, when both clustering and overdispersion are present. The subscript CO in $V_{CO}(\hat{\beta}_1)$ represents for clustered count data with overdispersion. We consider two overdispersion parameters: $\phi_0=1+\tau e^{\beta_0+\sigma^2}$ for unexposed and $\phi_1=1+\tau e^{\beta_0+\beta_1+\sigma^2}$ for exposed. We propose the following approximation of $\mathrm{Var}(\hat{\beta}_1)$ for clustered count data with overdispersion,

$$\mathrm{Var}(\hat{\beta}_1)=V_{CO}(\hat{\beta}_1)\cong\frac{1}{Nn}\left[\frac{\phi_0}{(1-R)\mu_0}+\frac{\phi_1}{R\mu_1}\right] \qquad (10)$$

where μ_0 and μ_1 are the same as those given in equation (5).

3. SIMULATIONS AND RESULTS

Under constant exposure proportion across clusters we simulated clustered Poisson data and clustered count data with overdispersion, and then evaluated the performance of the formulas in Sections 2.3 and 2.4.2 for sample size and power calculation. We also tested the robustness of these formulas in the presence of different exposure proportions for 10%, 20% and 30% of clusters.

3.1. Simulations

3.1.1. Clustered Poisson Data

We simulated clustered Poisson data under model (4) assuming $\gamma_i \sim N(0,\sigma^2)$. Because the unexposed group incidence rate in clustered Poisson data is associated with statistical power, in order to make valid comparisons of type I error rate and empirical power, we organized the simulation by different unexposed group incidence rates ($e^{\beta_0+\sigma^2/2}$). Thus in simulations we specified unexposed group incidence rate, and varied β_0 and σ^2 to achieve the specified rate. In detail, we specified the unexposed group log incidence rate ($\beta_0+\sigma^2/2$)to be 0.5, 0.8,1.0, or 1.5, and then varied β_0 from0.2 to 1.25 and σ^2 from 0 to 1.5 to obtain these incidence rates. This approach allows us to focus on the primary driver of power, unexposed incidence rate, while also examining any smaller effects of β_0 and σ^2 separately, on power. This approach is practically useful also, since incidence rates are typically better known by investigators than are the specific parameters β_0 and σ^2. We used two exposure proportions 0.5 (matching ratio=1:1, n=2) and 0.33 (matching ratio=1:2, n=3), and β_1=0.25 and 0.4. Given these parameters and target power=90%, sample sizes for unexposed and exposed subjects (n_0 and n_1) were calculated using the power formula (3) and $\mathrm{Var}(\hat{\beta}_1)$ in (5). For each combination of β_0,β_1,σ^2, and exposure proportion, 3000 clustered Poisson datasets with sample sizes n_0 and n_1 were simulated. Each simulated dataset was analyzed using SAS PROC NLMIXED (SAS Institute Inc., Cary, NC, v9.2) to fit model (4) for clustered Poisson regression and empirical power was calculated as the percentage of datasets with p<0.05 against the null hypothesis (β_1=0) based on the Wald test statistic. We also set β_1=0 to evaluate type I error rates.

To evaluate the robustness of the formula to the assumption of constant exposure proportion, we first calculated the required number of clusters as described above under a fixed exposure proportion (i.e., 0.33 (matching ratio=1:2)). We then allowed a portion of clusters to have different exposure proportions (i.e., 0.5 (matching ratio=1:1)). As a result, the sample size decreased slightly.

3.1.2. Clustered Count Data with Overdispersion

To simulate clustered count data with overdispersion under model (8), we used the same approach and same parameter values as above for simulating clustered Poisson data, with the addition of overdispersion using θ_{ij} ~ gamma ($1/\tau$, τ) with τ equal to 1.0 or 2.0. Given the above parameters and target power 90%, n_0 and n_1 were calculated based on the Wald test statistic using the power formula (3) and

$\mathrm{Var}(\hat{\beta}_1)$ in (10). Using the calculated n_0 and n_1, and each combination of the parameter values, 3000 datasets of clustered count data with overdispersion were simulated. Each simulated dataset was analyzed by maximizing the full likelihood function from model (8). This was accomplished in SAS PROC NLMIXED using the *general(ll)* statement by specifying a negative binomial conditional density and a cluster specific normal random effect. Empirical power was calculated as the percentage of datasets rejecting the null hypotheses (p<0.05). We also set $\beta_1 = 0$ to evaluate type I error rates. In addition, we examined the robustness of the power formula to the constant exposure proportion assumption as above.

3.2. Simulation Results

We evaluated our approaches by comparing type I error rates to nominal value 0.05 and by comparing the empirical powers to the pre-chosen target powers under which the sample sizes were determined and the count data were simulated.

3.2.1. Clustered Poisson Data

Our simulation results showed that estimates of β_1 were unbiased under the null and alternative hypotheses (data not shown). Type I error rates, under the null hypothesis β_1=0, were near 0.05 ranging from 0.04 to 0.06 under different intercept (β_0), exposure proportions, and variance of random cluster effects (σ^2) (data not shown).

Empirical power (Table **1**) was close to target power under $\beta_1 = 0.25$, different values of β_0, variance of random cluster effects (σ^2), and exposure proportions. For a fixed β_1, power is driven mainly by baseline incidence rate exp($\beta_0 + \sigma^2$/2), as shown by the large increases in required numbers of clusters as

Table 1: Empirical Power Obtained from 3000 Simulated Clustered Poisson Datasets with Target Power=90%, Constant Exposure Proportions 0.33 (Matching Ratio=1:2) or 0.5 (Matching Ratio=1:1), and β_1=0.25

$\beta_0 + \sigma^2/2$*	β_0	σ^2	Exposure proportion (exposed: unexposed)	Number of clusters (N)	Total number of subjects (Nn)	Empirical power (%)
1.5	1.25	0.5	0.5 (1:1)	67	134	89.7
	1	1				89.2
	0.75	1.5				86.9
	1.25	0.5	0.3 (1:2)	48	144	88.1
	1	1				87.3
	0.75	1.5				86.7
1.0	0.8	0.4	0.5 (1:1)	111	222	90.5
	0.5	1.0				88.5
	0.25	1.5				88.7
	0.8	0.4	0.3 (1:2)	79	237	88.6
	0.5	1.0				87.6
	0.25	1.5				87.5
0.8	0.8	0	0.5 (1:1)	135	270	90.4
	0.5	0.6				90.5
	0.2	1.2				89.2
	0.8	0	0.3 (1:2)	97	291	89.7
	0.5	0.6				88.5
	0.2	1.2				88.2
0.5	0.5	0	0.5 (1:1)	182	364	90.4
	0.2	0.6				90.6
	0.0	1.0				89.3
	0.5	0	0.3 (1:2)	130	390	89.1
	0.2	0.6				88.6
	0.0	1.0				88.4

*exp(β_0 +σ^2/2) is the mean incidence rate for unexposed subjects.

unexposed incidence rate decreases. This is consistent with the asymptotic variance in (5), which in addition to sample sizes and exposure effect depends only on this quantity. Lower incidence rates provide fewer events and less information to estimate the exposure effect. For a given baseline incidence rate (column 1), power decreased only slightly when σ^2 increased and β_0 decreased to achieve the same unexposed incidence rate. However the influences of both β_0 and σ^2 were small and empirical power remained above 86% within the range of parameters we tested. Note that these variances of random cluster effects represent very large between cluster variation; for σ^2=1.5, the 95% range of cluster-specific incidence rates spans a factor of $e^{1.96\sqrt{1.5}=11}$ above and below the cluster-specific incidence rate of $e^{\beta_0+\beta_1 x}$ for an average cluster (one with γ_i = 0). These results suggest that our newly proposed sample size approach performs reasonably well for clustered Poisson data. Similar results were obtained for exposure proportion 0.5 (1:1) and $\beta_1 = 0.4$ (data not shown).

Table **2** shows statistical power when $\beta_1 = 0.25$, exposure proportion for most clusters was 0.33 (matching ratio=1:2, n=3) and there was some variation in exposure proportion across clusters. When the majority of clusters had exposure proportion 0.33 (matching ratio=1:2) and a proportion of clusters ranging from 10% to 30% had exposure proportion 0.5 (matching ratio=1:1) (and thus the sample size dropped accordingly), empirical power dropped, but still remained above 83% for parameters tested. These results indicate that the formulas for sample size calculation for clustered Poisson data are reasonably robust to moderate deviation from the assumption of constant exposure proportion.

3.2.2. Clustered Count Data with Overdispersion

Our simulation results again showed that estimates of β_1 were unbiased under the null and alternative hypotheses (data not shown). Type I error rates, under the null hypothesis β_1 =0, were near 0.05 ranging from 0.04 to 0.06 under different exposure proportions, intercepts (β_0), variance of random cluster effects (σ^2), and subject heterogeneity (τ) (data not shown).

Table **3** shows empirical power for clustered count data with overdispersion for exposure proportion 0.33 (matching ratio=1:2) and $\beta_1 = 0.25$. As expected, the required number of clusters increases with increased overdispersion (τ). Patterns of sample size across individual parameters are more complex than for clustered Poisson data, but across the wide range of conditions we considered empirical power ranged from 89.2% to 95.4% for all combinations of parameters tested with target power 90%. Similar results were obtained for exposure proportion 0.5 (matching ratio=1:1) and $\beta_1 = 0.4$ (data not shown). These results suggest that the sample size calculation based on the newly proposed method performs well across a variety

Table 2: Empirical Power Obtained from 3000 Simulated Clustered Poisson Datasets with Target Power=90% and Constant Exposure Proportion=0.33 (Matching Ratio=1:2), $\beta_1 = 0.25$ and with Mixed Exposure Proportions

β_0 + σ^2 /2*	β_0	σ^2	Number of clusters (N)**	Empirical power (%)			
				All clusters 0.33 (1:2)	90% 1:2 10% 1:1	80% 1:2 20% 1:1	70% 1:2 30% 1:1
1.5	1.25	0.5	48	88.1	86.7	86.2	85.8
	1	1		87.3	86.2	86.0	83.9
	0.75	1.5		86.7	85.5	83.9	83.3
1.0	0.8	0.4	79	88.6	87.5	87.1	86.6
	0.5	1.0		87.6	86.6	86.5	85.3
	0.25	1.5		87.5	87.4	86.4	86.1
0.8	0.8	0	97	89.7	89.3	88.6	86.1
	0.5	0.6		88.5	87.6	87.8	87.2
	0.2	1.2		88.2	87.3	87.5	85.3
0.5	0.5	0	130	89.1	88.3	87.6	87.6
	0.2	0.6		88.6	88.0	85.9	85.7
	0	1		88.4	87.0	87.3	86.3

∗exp(β_0 +σ^2/2) is the mean incidence rate for unexposed subjects.

**There are 3 subjects in each cluster (one exposed and 2 unexposed). Thus the total sample size=3*number of clusters.

Table 3: Empirical Power Obtained from 3000 Clustered Count Datasets with Overdispersion when Target Power=90% and $\beta_1 = 0.25$, with Constant Exposure Proportion 0.33 (Matching Ratio=1:2), and with Mixed Exposure Proportions

					Empirical powers (%) under mixed exposure proportions for clusters			
τ	$\beta_0 + \sigma^2/2$*	β_0	σ^2	**Number of clusters (N)****	**All clusters 0.33 (1:2)**	**90% 1:2 10% 1:1**	**80% 1:2 20% 1:1**	**70% 1:2 30% 1:1**
1	1.5	1.25	0.5	372	91.2	90.2	90.2	89.3
		1	1.0	464	92.2	92.5	91.6	91.4
		0.8	1.4	556	95.4	95.4	94.0	93.6
	1.0	0.8	0.4	387	90.2	88.6	87.5	87.3
		0.5	1.0	495	91.4	90.6	90.3	89.6
		0.25	1.5	613	94.8	94.1	93.8	93.3
	0.8	0.8	0	348	90.1	90.2	89.3	88.5
		0.5	0.6	437	90.2	90.4	89.1	87.1
		0.2	1.2	556	91.6	90.8	90.4	89.4
2	1.5	1.25	0.5	695	91.2	90.7	89.4	88.9
		1.0	1.0	878	92.9	92.2	91.8	91.3
		0.8	1.4	1063	94.8	94.6	93.9	93.8
	1.0	0.8	0.4	695	90.0	88.2	88.2	87.3
		0.5	1	910	92.6	92.2	91.3	90.4
		0.25	1.5	1146	95.1	94.5	94.1	93.7
	0.8	0.8	0	601	89.5	89.4	89.3	88.3
		0.5	0.6	777	90.8	90.3	90.1	88.1
		0.2	1.2	1015	92.4	91.6	91.7	90.9
	0.5	0.5	0	634	89.3	90.2	89.2	88.5
		0.2	0.6	811	90.2	89.9	89.5	88.3
		0	1	962	91.4	89.7	89.7	88.9

∗exp(β_0 +σ^2/2) is the mean incidence rate for unexposed subjects; **There are 3 subjects in each cluster (one exposed and 2 unexposed). Thus for the balanced cluster case the total sample size=3*number of clusters, and is reduced accordingly when unbalanced clusters are considered.

of conditions for clustered count data with overdispersion.

In Table **3** we also show empirical power when the majority of clusters had 0.33 exposure proportion (matching ratio=1:2) and a proportion of clusters ranging from 10% to 30% had 0.5 exposure proportion (matching ratio=1:1). Empirical power changed slightly with varying exposure proportions but remained above 87% for all parameter combinations considered despite the loss of some subjects and reduction of total sample size. These results indicate that the formulas for sample size estimation continue to be robust to moderate deviation from the assumption of constant exposure proportion for clustered count data with overdispersion.

4. EXAMPLE: ASSOCIATION BETWEEN PARENTAL DEPRESSION AND CHILDREN'S HEALTH CARE USE

The Kaiser Permanente Colorado study of health care use by children exposed to parents with a diagnosis of depression was described briefly in the Introduction. We compared the incidence of total clinical visits between exposed and unexposed children. There were a total of 26,104 matched clusters, of which 20,883 (80.0%) had 0.33 exposure proportion (matching ratio=1:2). The remaining clusters had 0.5 exposure proportion (matching ratio=1:1). The mean and variance of the total number of clinical visits were 1.50 and 5.34, respectively, indicating the presence of overdispersion. These matched cohort data were analyzed by fitting model (8) using SAS

NLMIXED. The parameter estimates are $\hat{\beta}_0$=0.04 (intercept), $\hat{\beta}_1$=0.45 with 95% CIs (0.42, 0.47) (exposure effect), $\hat{\sigma}^2$=0.4 (variance of random cluster effects), and $\hat{\tau} = 0.77$ (subject heterogeneity). Both clustering and overdispersion are present.

Assuming exposure effect size $\hat{\beta}_1$=0.45, based on the sample size/power formulas in Section 2.4.2, we would have needed only 91 and 120 matched clusters (matching ratio=1:2) to achieve 80% and 90% power, respectively. We verified the power calculations using a bootstrap approach with the following steps: 1) randomly select 91 or 120 clusters out of the 20,883 available clusters with matching ratio=1:2; 2) analyzed the random sample using model (8) with NLMIXED; 3) repeat steps 1 and 2 1000 times. Empirical powers were calculated as the proportion of bootstrap datasets with p-value for the exposure effect less than 0.05. The empirical powers were 79% and 88% for sample sizes based on true powers 80% and 90%, respectively.

We now show how the methods we have provided, along with information from the Kaiser study, would be used in designing a future study. We consider detecting an increase in incidence rate of 30% for total clinic visits of children exposed to a parent diagnosed with depression, i.e. incidence rate ratio=1.30 or β_1= log(1.3)=0.262. We assume a two-sided level 0.05 test with 90% power, and an incidence rate for children not exposed to a parent diagnosed with depression similar to that noted in the previous study, exp($\hat{\beta}_0 + \hat{\sigma}^2$/2) =exp(0.04 + 0.4/2) = 1.27. We further assume two unexposed children for each exposed child. If only clustering was considered with σ^2= 0.40, the required sample sizes using equations (3) and (5) are n_1= 153 and n_0= 306, or 153 matched clusters of one exposed and two unexposed children; if both clustering and overdispersion were considered with τ = 0.77 and σ^2= 0.40, the required sample sizes using equations (3) and (10) are n_1= 369 and n_0= 738, or 369 clusters of one exposed and two unexposed children. Accounting for overdispersion the required number of clusters increased dramatically.

5. DISCUSSIONS

In this paper we provided simple formulas for sample size and power calculation for matched cohort study designs when the outcome is a count and the exposure proportion is constant or nearly constant across clusters. We considered both clustered Poisson data, which has previously been discussed [8, 28], and clustered count data with overdispersion, which to our knowledge has not been previously considered. We derived theoretical expressions for the variance of the estimated exposure effect for the clustered Poisson case, and provided an approximation for the variance of the estimated exposure effect for the clustered count data with overdispersion. Simulation studies showed that the sample size formula based on a Wald test statistic yielded robust type I error rate (0.04-0.06) under the null hypothesis and gave empirical power close to the target power across a range of parameter values for both clustered Poisson data and clustered count data with overdispersion. In relation to previous work, we noticed that for clustered Poisson data, our calculated power is slightly lower than the targeted power, yet it was slightly higher than the empirical power in Amatya [8] although we used the same formula for $\mathrm{Var}(\hat{\beta}_1)$. The discrepancy could be due to two factors. First, we used $\mathrm{Var}(\hat{\beta}_1)$ under the alternative hypothesis for our test statistic, while Amatya used $\mathrm{Var}(\hat{\beta}_1)$ under both null and alternative hypotheses for their test statistic. Second, the selected baseline incidence rates were much lower in their Table **1** than those in our Table **1**.

Simulations also showed that these sample size and power estimates are robust to moderate deviations from the assumption of constant exposure proportion. Inclusion of up to 30% of clusters with a different exposure proportion only slightly reduced empirical power (less than 4%). This is important because practically studies will typically be designed with constant exposure proportion, but during implementation the final achieved exposure proportions may not be constant due to unavailability of some subjects.

The number of clusters required to achieve a target power increased dramatically for clustered count data with overdispersion compared with clustered Poisson data (Tables **1** and **3**). This emphasizes the importance of incorporating overdispersion during the planning phase. In real count data the conditionally Poisson assumption is typically not satisfied, and overdispersion tends to be the norm.

Although most of the times the empirical power is lower than the true (90%) in Table **1** for clustered Poisson data, the majority is within 1% and the greatest difference from the true is 3.3%. The empirical power is almost always greater than the true (90%) in Table **3** for clustered Poisson data with overdispersion, the majority is within 2% and the greatest difference from the true is 5.1%. We suspect the overestimation of sample size thus the empirical power is due to slightly overestimation of $\mathrm{Var}(\hat{\beta}_1)$ for clustered count data with overdispersion. This phenomenon warranties future exploration of a correcting method.

Our results apply to a within cluster treatment or covariate such as would occur when clusters are identified and some subjects within each cluster receive the treatment, or when subjects with and without a covariate value are identified from a database and matched on other characteristics. These results complement the situation of cluster randomized designs where all subjects in a cluster receive the same treatment or covariate level. In the latter case the results of Gao [14] and Demidenko [15] do not apply, and the standard methods for cluster randomized designs [27, 29, 30] can be used.

For clarity of presentation we have discussed subjects within clusters, but these methods would also apply to other forms of clustering, for example a count outcome measured on subjects while on a treatment and while off the treatment, as in a time series intervention design. Here, subjects are clusters and treatment is a within cluster variable.

The reasons for us to consider only random cluster effects models rather than a general population-average model (i.e., Generalized Estimating Equations (GEE)) are two-fold: first, it has been shown that the results from GEEs and random-effects model were comparable in analyzing clustered count data [31]. Our preliminary analyses also showed that the empirical power based on GEEs was very close to the empirical power based on random cluster effects model. Second, it is advantageous to use random cluster effects model in deriving the sample size and power formulas because of its explicit definition of the model and available moment-generating functions.

ACKNOWLEDGEMENT

Gao and Xu were partially supported by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780.

CONFLICTS OF INTEREST

All authors have no conflicts of interest.

REFERENCES

[1] Kornek B, Aboul-Enein F, Rostasy K, *et al*. Natalizumab therapy for highly active pediatric multiple sclerosis. JAMA Neurol 2013; 70: 469-75.

[2] Rothman KJ, Greenland S. Cohort Studies. In Modern Epidemiology, 2nd edition, Philadelphia, PA: Lippincott-Raven 1998.

[3] Graham PL, Mengersen K, Morton AP. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. Statistics in Medicine 2003; 22: 2071-2083. http://dx.doi.org/10.1002/sim.1405

[4] Cummings P, McKnight B, Greenland S. Matched cohort methods in injury research. Epidemiologic Reviews 2003; 25: 43-50. http://dx.doi.org/10.1093/epirev/mxg002

[5] Cummings P, McKnight B, Weiss NS. Matched-pair cohort methods in traffic crash research. Accident Analysis and Prevention 2003; 35: 131-141. http://dx.doi.org/10.1016/S0001-4575(01)00108-7

[6] Sills MR, Shetterly S, Xu S, Magid D, Kempe A. The association between parental depression and children's healthcare utilization. Pediatrics 2007; 119: e829-836.

[7] Ng HKT, Tang ML. Testing the equality of two Poisson means using the rate ratio. Statistics in Medicine 2005; 24: 955-965. http://dx.doi.org/10.1002/sim.1949

[8] Amatya A, Bhaumik D, Gibbons RD. Sample size determination for clustered count data..Statistics in Medicine 2013; 32: 4162-4179.

[9] Cox DR. Some remarks on overdispersion. Biometrics 1983; 10: 269-274.

[10] Dean C. Testing for overdispersion in Poisson and binomial regression models. Journal of the American Statistical Association 1992; 87: 451-457. http://dx.doi.org/10.1080/01621459.1992.10475225

[11] Lawless JF. Negative Binomial and Mixed Poisson Regression. The Canadian Journal of Statistics 1987; 15: 209-225. http://dx.doi.org/10.2307/3314912

[12] Cameron AC, Trivedi PK. Regression Analysis of Count Data. Cambridge University Press 1998. http://dx.doi.org/10.1017/CBO9780511814365

[13] Friede T, Schmidli H. Blinded sample size re-estimation with count data: Methods and applications in multiple sclerosis. Statistics in Medicine 2010; 29: 1145-1156.

[14] Gao D. Analysis of clustered longitudinal count data. University of Colorado Health Sciences Center Thesis 2007.

[15] Demidenko E. Poisson regression for clustered data. International Statistical Review 2007; 75: 96-113. http://dx.doi.org/10.1111/j.1751-5823.2006.00003.x

[16] Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of Longitudinal Data. 2nd ed. Oxford University Press: New York 2002.

[17] Breslow N. Test of hypotheses in overdispersion regression and other quasi likelihood models. Journal of the American Statistical Association 1990; 85: 565-571. http://dx.doi.org/10.1080/01621459.1990.10476236

[18] Nagin DS, Land KC. Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model. Criminology 1993; 31: 501-523. http://dx.doi.org/10.1111/j.1745-9125.1993.tb01133.x

[19] Nagin DS. Group-Based Modeling of Development. Cambridge: Harvard University Press 2005.

[20] Sichel HS. The density and size distribution of diamonds. Bulletin of the International Statistical Institute 1973; 45: 420–427.

[21] Atkinson AC, Yeh L. Inference for Sichel's compound Poisson distribution. Journal of the American Statistical Association 1982; 77: 153-158. http://dx.doi.org/10.1080/01621459.1982.10477779

[22] Manton KG, Woodbury MA, Stallard E. A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients

in lung cancer mortality rates in north Carolina counties. Biometrics 1981; 37: 259-269. http://dx.doi.org/10.2307/2530416

[23] Margolin BH, Kaplan N, Zeiger E. Statistical analysis of the Ames Salmonella Microsome Test. Proceedings of the National Academy of Sciences 1981; 76: 3779-3783. http://dx.doi.org/10.1073/pnas.78.6.3779

[24] Hinde J. Compound Poisson regression models. Lecture Notes in Statistics 1982; 14: 109-121. http://dx.doi.org/10.1007/978-1-4612-5771-4_11

[25] Ord JK, Whitmore GA. The Poisson-inverse Gaussian distribution as a model for species abundance. Communications in Statistics-Theory and Methods 1986; 15: 853-871. http://dx.doi.org/10.1080/03610928608829156

[26] Hougaard P, Lee MLT, Whitmore GA. Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. Biometrics 1997; 53: 1225-1238. http://dx.doi.org/10.2307/2533492

[27] Molenberghs G, Verbeke G, Demétrio CGB. An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis 2007; 13: 513-531. http://dx.doi.org/10.1007/s10985-007-9064-y

[28] Ogungbenro K, Aarons L. Sample size/power calculations for population pharmacodynamic experiments involving repeated-count measurements. Journal of Biopharmaceutical Statistics 2010; 20: 1026-1042. http://dx.doi.org/10.1080/10543401003619205

[29] Cornfield J. Randomization by group: a formal analysis. American Journal of Epidemiology 1978; 108: 100-102.

[30] Donner A, Klar N. Design and analysis of cluster randomization trials in health research. Arnold: London; 2000.

[31] Gao D, Grunwald G, Xu S. Statistical Methods for Estimating Within-Cluster Effects for Clustered Poisson Data. J Biomet Biostat 2013; 4: 1-6.

Received on 07-07-2014 Accepted on 04-08-2014 Published on 25-08-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.03.11