Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 263–274

Estimation and Sample Size Calculation for Service Utilization Data

Dulal K. Bhaumik[1] **and Subhash Aryal**[2]
[1]*Department of Psychiatry, University of Illinois at Chicago, Chicago, USA*
[2]*School of Nursing, University of Pennsylvania, Philadelphia, USA*

Received: 01 June 2020; Revised: 29 July 2020; Accepted: 06 August 2020

Abstract

Health service utilization research suffers from lack of statistical methods to analyze routinely obtained zero-inflated correlated outcome data from multilevel longitudinal studies. Parameter estimation suffers from use of maximum likelihood based approach involving cumbersome integration which results in lack of model convergence and utilization of considerable computing resources. Similarly, sample size to conduct randomized controlled trials are estimated using either inappropriate linear models or simplified non-linear models which ignore multiple levels of nesting resulting in severely under powered studies. We propose a robust estimation method based upon Laplace approximation to estimate parameters and derive formula to compute required sample size employing multiple levels of nesting.

Key words: Health services; Zero-inflated data; Laplace approximation; Sample size.

1. Introduction

Health services (HS) researchers are widely using hierarchical mixed-effects models for analysis of their correlated clustered and longitudinal data. Parameters are generally estimated by maximum marginal likelihood, empirical Bayes estimation, fully Bayesian strategies and Generalized Estimating Equations (GEE), and hypotheses are tested using t, χ^2 or F tests. Furthermore, considerable computer software has now been developed and is either freely available over the Internet or commercially available. However, this area is still challenged by a lack of statistical methods appropriate for addressing some unique aspects of health services research data. A major problem in HS data is missing outcomes as well as covariate values. Another equally complex problem is the profusion of zero values in count data such as service units or costs, which results in a highly skewed distribution. To address these issues, in many instances missing values are imputed, and hierarchical zero-inflated mixed models are utilized even though non-convergence issues prevail in estimation. In such models, justification of using random effects in terms of testing its variance components is avoided because of unavailability of user-friendly testing procedures at the boundary value. Another challenge is the determination of sample size, as inadequate sample size runs the

Correponding Author: Dulal K. Bhaumik
Email: dbhaumik@uic.edu

risk of inflated false positive findings (Type I error), while fitting the model with an excessive number of random-effects can mask significant relationships (Type II error). This manuscript addresses issues pertaining to parameter estimation and sample size calculation for HS researches and bridge a critically important gap in the designing stage of health research studies in general, and mental health services research in particular.

In *mental health services research* investigators have studied service utilization, barriers to service utilization, disparities in service utilization and cost associated with service utilization (Hacker, *et al.* 2015). Similarly, service utilization data are also found in research on *general health care* (Gilbert, *et al.* 2012), *dentistry* (Moghimbeigi, *et al.* 2008), *occupational health* (Min and Agresti, 2005) and *substance abuse* (Bandhophadyay, *et al.* 2011). Our careful analysis of the literature revealed that service utilization research studies regularly encounter the problem of missing outcomes and covariates, zero-inflation, over-dispersion, and non-convergence of statistical models. In addition, this area requires feasible parameter estimation techniques and sample size determination methods and user friendly software for analysis of HS data. Particularly there is an lack of suitable software for sample size determination when zero-inflation is expected in a hierarchical design with random-effects. Most of the existing methods either assume linear model or completely ignore the random-effects by using the GEE approach. As such there is a genuine need for sample size methodologies and more importantly software to calculate sample size for service utilization research with zero-inflation.

In Section 2, we present some motivating examples. In Section 3, we discuss methods to model service utilization data. In Section 4, we derive formulae for sample size calculation for studies employing hierarchical designs resulting in zero-inflated outcomes . In Section 5, we present some concluding remarks on service utilization data.

2. Motivating Examples

Next we present two HS research studies to motivate the need for theoretical developments.

2.1. Example 1

The first problem was investigated by Atkins, *et al.* (2015) and compared group differences between Links to Learning (L2L), a school and home-based mental health service model, and Service As Usual (SAU) on several domains including mental health service use, classroom observations of academic engagement, teacher report of academic competence and social skills, parent report of social skills, teacher and parent report of problem behaviors, daily hassles, and curriculum-based measures. Services were Medicaid-funded through 4 social service agencies ($N = 17$ providers) in 7 schools ($N = 136$ teachers, 171 children consists of 124 boys (50 control + 74 Link), and 47 girls (17 control + 30 Link)) in a 2 (Links to Learning vs. services as usual) 6 (pre- and post tests for 3 years) longitudinal design with random assignment of schools to conditions. Services as usual consisted of supported referral to a nearby social service agency. The primary interest was in *differential change over time.* A three-level hierarchical design with multiple observations from students nested within schools was used to analyze the study data. The model included covariates at both the student level (grade, gender) and classroom and teacher level (classroom assessment scoring system, teacher sense of efficacy scale, organizational health inventory-elementary and Quality of teacher work life survey). The conclusion of the study was that community mental health services targeting empirical predictors of learning can improve school and home behavior for children living in high-poverty urban communities. For a full description of data decomposition, analysis methods, missing value problems and significant results we refer to Atkins, *et al.* (2015). Some key difficulties encountered during the analysis of this dataset were (i) problem of missing data (more than 41.79% in control and 52.88% in L2L data), (ii) differential measurement errors, and (iii) the problem of unreliable measures of some outcomes and covariates. This analysis inspired us to develop novel statistical methods to estimate missing outcomes when corresponding covariates are known, and missing covariates when corresponding outcome measures are known, but in both situations causes of missingness are unknown.

2.2. Example 2

Our second example is based on the work by Cook, *et al.* (2019) and Bhaumik, *et al.* (2019). They recently analyzed data from randomized trial of self-directed care in Texas public mental health system. In this study, the Zero-Inflated Negative Binomial (ZINB) and log-gamma models were used to test the effect of an experimental intervention called self-directed care, in which patients have greater control over service delivery funds and can choose to hire and fire specific service providers. The authors applied the ZINB model to analyze service utilization and log-gamma model for analysis of cost data. A total of 216 subjects with serious mental illness receiving care in the Texas public mental health system were randomly assigned with their consent to receive services as usual (= 102) or the experimental intervention (= 114) and followed for 24 months. The primary hypothesis was that the experimental intervention would produce superior client outcomes at 12 and 24-month follow-up and this proved to be the case. However, since the intervention was intended to be budget neutral (*i.e.*, to cost no more than services delivered through the usual system), secondary analysis of service costs was required. Administrative data were obtained from the local area's managed care company in the form of "shadow claims" and grouped into costs during the first and second years of program participation and for both years combined. Over the two years of the program, experimental participants incurred a total average per person cost of $5,239(s.d. = 5,500)$ compared to an average of $5,493(s.d. = 8,268)$ per person in the control group. This difference was non-significant, as expected. However, costs for specific service types had the additional challenge of being zero-inflated, with many non-users of some services. Consequently, the authors used ZINB/log-gamma models for individual services/costs, which model the mixture of the likelihood of having zero service/costs in each category, and the relative amount of service/costs among users. As shown in Table 1, experimental condition subjects were more likely than controls to have zero costs for psychiatric rehabilitation, case management, and skills training, but there were no differences in costs for users of these three services. On the other hand, there was no difference in the likelihood of zero costs for medication management, but among users of this service, costs were significantly lower for the experimental group. For the service of psychotherapy, the experimental group was less likely than controls to have zero costs, and costs were higher for experimental than control subjects. When the authors used linear mixed-effects regression analysis of these individual service costs adjusting for time, the experimental condition costs were lower for psychiatric rehabilitation, skills training, and medication management, and higher for psychotherapy. The linear mixed-effects model cannot provide information separately for zero costs and costs for users. Clearly, ZINB modeling provided a more complex and complete picture of cost differences where they existed.

Table 1: ZINB analysis modeling first, likelihood of zero costs, and second, costs among service users

	Psychiatric Rehab		Case Mgmt		Skills Training		Medication Mgmt		Psycho-therapy	
	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*	Estimate	*p*
Pred. zero cost	0.755	.007	1.183	.001	1.484	<.001	0.731	.076	-1.602	<.001
Pred. costs users	-0.297	.124	0.076	.855	-0.490	.151	-0.439	.001	1.134	.001

There are important implications in these results for health services researchers and the service system administrators and policy makers who use these study findings. First, ZINB allowed us to use a "two-part model" in analyzing utilization data (Manning, *et al.* 2005). As noted by Diehr, *et al.* (1999), the decision to have any use of a service is most likely made by the person and so is primarily associated with personal characteristics, while the cost per user may be more related to features of the health care system. When the goal is understanding the system, a two-part model is preferred because it enables researchers to differentiate between influences on the propensity to use a service, and factors affecting how much of the service is used and at what cost once the individual enters the service delivery system (Diehr, *et al.* 1999).

3. Model

In this section we present some models for analysis of count data inflated with zeros. We begin by positing c centers and n_i subjects nested within the ith center. The total number of subjects $N = \sum_{i=1}^{c} n_i$ are randomized into intervention and control groups. We assume that each subject may utilize mental health services longitudinally over T different time periods. The outcome variable y_{isjt} measures the number of times the jth subject from the ith center nested within the sth intervention group used mental health services for the tth time period. The log likelihood function for all observations $\boldsymbol{y}_i = (y_{i111}, \cdots, y_{i2nT})^t$ nested within the ith center is

$$logit(\pi_{isjt}) = \gamma_0 + \gamma_1 g(t) + \gamma_2 x_{ijk} + \gamma_3 x_{ijk} g(t) + \boldsymbol{\gamma}^{t*} \boldsymbol{w_{ijk}} + \nu_{i0} + \nu_{i1} g(t) + \delta_{isj0} + \delta_{isj1} g(t), \quad (1)$$

$$log(\lambda_{isjt}) = \beta_0 + \beta_1 g(t) + \beta_2 x_{ijk} + \beta_3 x_{ijk} g(t) + \boldsymbol{\beta}^{t*} \boldsymbol{z_{ijk}} + \upsilon_{i0} + \upsilon_{i1} g(t) + \delta^*_{isj0} + \delta^*_{isj1} g(t), \quad (2)$$

Here $f(y_{isjt})$ be the probability mass function of a Poisson distribution for a zero inflated Poisson (ZIP) model, and if the model is zero inflated negative binomial then $f(y_{isjt} = k) = (1 - \pi_{ij}) \frac{\Gamma(k + \lambda_{ij}^{1-d}/\alpha)}{k\Gamma(-\lambda_{ij}^{1-d}/\alpha)} (1 + \alpha\lambda_{ij}^{d})^{-\lambda_{ij}^{1-d}/\alpha} (1 + \lambda_{ij}^{-d}/\alpha)^{-k}$. The dispersion parameter

α is always non-negative and does not depend on covariates. This distribution reduces to the ZIP distribution when $\alpha \to 0$. The constant d is used to identify a particular form of a negative binomial distribution (see Saha and Dong, 1997). For various forms of negative binomial distributions we refer to (McCullagh and Nelder 1989, Ridout, *et al.* 2001, Yau, *et al.* 2003). Xiang, *et al.* (2007) used a score test for testing the over-dispersion of a ZIP regression model against the ZINB alternative (*i.e.*, $\alpha = 0$ in a ZINB model). Both ZIP and ZINB regression models will be inappropriate for fitting data with zero deflation at any settings of the explanatory variables. A useful model for such a situation is the "Hurdle" model proposed first by Mullahy (1986) that separately handles the zero observations and the positive counts. An advantage of the Hurdle model is that it can handle both the zero inflation and zero deflation. The downside of this model is that all zero counts are structural whereas ZINB and ZIP models allow both structural and functional zeros (Pardoe and Durham, 2003). In the Hurdle model, $g_2(y_{ij}) = f(k)/(1 - f(0))$.

In model (1), $\gamma_0 + \gamma_1 g(t)$ and $\gamma_0 + \gamma_1 g(t) + \gamma_2 x_{ijk} + \gamma_3 x_{ijk} g(t)$ are the fixed linear trends for the control group, and for the intervention group, respectively. Thus γ_3 differentiates the slope of the treatment group from the control group of service utilization and β_3 has a similar interpretation of frequency of service utilization. Exponentiation of γ_3 and β_3 provides the odds ratio and risk ratio respectively. Note that these parameter estimates are subject-specific, which indicates the effectiveness of the intervention at the individual level. The interpretation of $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ is of considerable interest. The gamma parameters ($\boldsymbol{\gamma}^*$) describe the effects of the covariates on the likelihood of service utilization, whereas the beta parameters ($\boldsymbol{\beta}^*$) describe the effects of the same or possibly different covariates on the intensity of service utilization. Also, $\nu_{i0} + \nu_{i1} g(t)$ is the random linear trend for the ith site effect. The correlation between subjects nested within the same site is accounted for by the presence of random site effects. Similarly $\delta_{isj0} + \delta_{isj1} g(t)$ is the random linear trend for the jth subject nested within the ith site, and the random linear trend at the subject level takes care of the correlation between multiple observations nested within the same subject. Similar interpretations hold in model (2). The vectors $\boldsymbol{w_{ijk}}$ and $\boldsymbol{z_{ijk}}$ represent the additional fixed covariates such as age, race, sex etc. for the logit and the log-linear components. A three-level ZIP or ZINB longitudinal mixed-effects model can have a total of 12 variance covariance parameters; six components from the binary part (variance for random intercept, variance for random slope and their covariance for subjects and for communities), and a parallel set of six variance components from the count part of the model. Even though 12 variance components in the above models seems to be a reasonable assumption, in actuality, we do not know how many of them are really significant. Keeping all of them may over-saturate the model. To select an appropriate model we generally use deviance, Akaike information criterion (AIC), and Bayesian information criterion (BIC). Several authors have recently noted that AIC and BIC are not appropriate for model selection when the sample size is small (Kass and Raftery, 1995, Seghouane, 2006, Chen, *et al.* 2008, Tu and Xu, 2012). To resolve this issue, there is a need for alternative approaches to evaluate the significance of variance components.

3.1. Estimation of model parameters

The goal of this section is to derive and use the marginal likelihood function of fixed parameters (*i.e.* γ and β) conditioning on the data and a suitable estimate of random effects (posterior mode). At the initial stage, it is assumed that variance components are known. The numerical integration over the space of random-effects in the estimation process is avoided by approximating the log-likelihood around the starting values of random-effects. In addition, we investigate an alternative procedure based on ordinary Laplace approximation. The convergence rates of both ordinary Laplace and Marginal Maximum Likelihood (MML) (combination of Gaussian quadrature and Newton Rhapson method) is $O(n^{-1})$ (Ghosh, *et al.* 2006, page 206). However, MML requires enormous computational time and often fails to converge for hierarchical Zero-Inflated Data (ZID). Xie, *et al.* (2013) and Gupta, *et al.* (2015) encountered similar convergence problem in their analysis of ZID. On the other hand, the Laplace approximation avoids numerical integration by exploiting a property of the multivariate normal distribution. As a result, this method provides better guarantee of convergence compared to the quadrature methods for hierarchical models. For comparison purposes we also include the penalized quasi likelihood (PQL) approach (Hyede, 1997).

3.2. Comparison of three estimation methods

First, zero-inflated data were simulated under the assumption that all random-effects in the logistic component were stochastically independent from the random-effects in the log linear component. This assumption reduced the complexity of numerical computation. To compare results of parameter estimation obtained by Laplace and quadrature methods for a two-level Poisson and logistic regression mixed-effects models, we set intercept parameters of control and intervention groups at 3 and 0, respectively, and slope parameters at -0.5 and -1, respectively. The variance-covariance matrix of the random slope and random intercept were set at (1, -0.2, 0.05). Based on simulations using a two-level ZI model, we observed in Table that Laplace and quadrature methods produced similar results, whereas results by PQL were unsatisfactory. In addition, we observed in Table that standard errors of these estimates for both Laplace and quadrature methods did not vary significantly. However, the convergence rate obtained by the Laplace method was substantially higher than that of the quadrature method. In addition, the Laplace method required, on average, one-fourth of the computing time required by the MML method (whenever it converged), and the accuracy rates of both Laplace and quadrature were at the same level.

3.3. Computation time and convergence

Another critically important issue in fitting complex models with numerous random-effects is computational time and model convergence. To investigate these issues we used PROC NLMIXED, SAS version 9.4 to fit our models. The computational times for ZIP and ZINB models (i) with fixed-effects, are in terms of seconds, (ii) with mixed-effects having one or two random-effects, are less than 5 minutes for both quadrature and Laplace, (iii) with mixed-effects having three random-effects, are around 80 minutes for quadrature and less than 25 minutes for Laplace, (iv) the quadrature did not converge for both ZIP and ZINB models with four random-effects, whereas, for the same models with four random effects, Laplace converged in two hours. The same data analyzed using GEE took less than

Table 2: Estimation of parameters and standard errors by Laplace, Quadrature and PQL for Poisson and logistic regression models.

	Laplace		Quadrature		PQL	
Parameters	Poisson	Logistic	Poisson	Logistic	Poisson	Logistic
$\beta_0 = \gamma_0 = 3$	2.997 (0.14)	3.31 (0.63)	2.997 (0.14)	3.202 (0.68)	3.022 (0.14)	2.736 (0.87)
$\beta_1 = \gamma_1 = -0.5$	-0.499 (0.04)	-0.558 (0.14)	-0.498 (0.04)	-0.537 (0.14)	-0.478 (0.04)	-0.463 (0.08)
$\beta_2 = \gamma_2 = 0$	0.007 (0.20)	-0.018 (0.64)	0.007 (0.20)	0.039 (0.62)	-0.016 (0.25)	-0.082 (0.72)
$\beta_3 = \gamma_3 = -0.5$	-0.503 (0.07)	-0.543 (0.21)	-0.504 (0.07)	-0.557 (0.25)	-0.490 (0.37)	-0.412 (0.54)
$\sigma^2_{\delta_0} = \sigma^2_{\delta^*_0} = 1$	0.979 (0.14)	1.412 (1.53)	0.981 (0.15)	1.606 (1.60)	0.959 (0.25)	0.219 (0.64)
$\sigma_{\delta_{01}} = \sigma_{\delta^*_{01}} = -0.2$	-0.197 (0.04)	-0.573 (0.53)	-0.197 (0.04)	-0.354 (0.40)	-0.204 (0.09)	-0.005 (0.79)
$\sigma^2_{\delta_1} = \sigma^2_{\delta^*_1} = 0.05$	0.086 (0.01)	0.149 (0.15)	0.086 (0.01)	0.154 (0.21)	0.083 (0.11)	0.012 (0.25)

one minute with an exchangeable correlation matrix. The Bayesian approach with three random-effects took 10 minutes to update 1 chain for 10,000 iterations (5000 burn-in, 5000 update), and 15 minutes to update 2 chains for 10,000 iterations (5000 burn-in, 5000 update). We further repeated the simulation study with various levels of missingness and observed that computational time varied significantly between the methods, and non-convergence became a norm rather than an exception, especially when missingness exceeded more than 30%. An alternative approach when convergence persists is the use of "Maximum A Posteriori (MAP)" estimation that sets the initial value of the parameters to their posterior mode, and uses adaptive quadrature instead of fixed-point quadrature. Yet another alternative is to use the Laplace approximation at each center, and then perform meta analysis to combine results from centers (Bhaumik, *et al.* 2012, Amatya, *et al.* 2015). Convergence rate for this combination approach is expected to be better as random components at the center level are eliminated. Based on this simulation study, we recommend to use Laplace method (or a combination of Laplace and meta analysis) for estimating parameters of zero inflated models when number of random effects is more than two in order to get consistent estimators avoiding non-convergence issues.

4. Sample Size Determination

In this section we address the issue of sample size determination for hierarchical designs with zero-inflated data.

Statistical methods for the analysis of longitudinal data with clustering of subjects are now routinely applied in mental health service utilization studies. The design of such studies often suffers from poorly specified and often inadequate sample sizes. This is because sample size determination methodology is derived based on a single outcome, or based on longitudinal studies which ignore clustering. The determination of sample sizes when subjects are both repeatedly measured over time and clustered within research sites (*e.g.*, multisite Randomized Controlled Trials (RCTs)) can be erroneous unless both factors, and attrition rates are taken into account. Several authors have developed power analysis for cluster-randomized, and/or repeated measurements studies (Roy, *et al.* 2007, Bhaumik, *et al.* 2008, 2013, Amatya, *et al.* 2013, Kapur, *et al.* 2014). Some of the key features of power calculations include (i) type of randomization (participant level, or site level), (ii) cluster and longitudinal variability, (iii) differential attrition rates over time, and also in different groups (*i.e.* intervention and control groups), and (iv) proportion of allocations of subjects.

4.1. Theoretical foundation for sample size computation using generalized linear models

Denote the outcome of the ith subject nested within the cth cluster measured at the jth time point by y_{cij}, where $i = 1, \ldots, n$, $c = 1, \ldots, C$, and $j = 1, \ldots, T$. Let $\boldsymbol{y}_{ci}$= $(y_{ci1}, y_{ci2}, \cdots, y_{ciT})^t$ be a column vector of dimension $T \times 1$ composed of outcomes of the ith subject measured at T different time points. Generalized linear mixed model that links the expectation of $\boldsymbol{y}_{ci}$ to the linear predictor has the following expression: $E(\boldsymbol{y}_{ci}|\boldsymbol{\delta}, \boldsymbol{\gamma}) = \mathbf{h}(\boldsymbol{X}_{ci}\boldsymbol{\beta} + \boldsymbol{Z}_{ci}\boldsymbol{\delta}_{ci} + \boldsymbol{W}_{ci}\boldsymbol{\gamma}_c) = \mathbf{h}(\boldsymbol{\eta}_{ci})$, where $\boldsymbol{\eta}_{ci} = \boldsymbol{X}_{ci}\boldsymbol{\beta} + \boldsymbol{Z}_{ci}\boldsymbol{\delta}_{ci} + \boldsymbol{W}_{ci}\boldsymbol{\gamma}_c$, and $\boldsymbol{X}_{ci}$, $\boldsymbol{Z}_{ci}$ and $\boldsymbol{W}_{ci}$ are design matrices associated with fixed-effects ($\boldsymbol{\beta}$), subject-level random-effects ($\boldsymbol{\delta}_{ci}$), and cluster-level random-effects ($\boldsymbol{\gamma}_c$), respectively. Random-effects $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are independent and assumed to follow multivariate distributions. Denote the number of clusters by C, number of treatments by S, the covariance matrix of the pseudo observation $\boldsymbol{y}^*$ (obtained by linearizing the real observation $\boldsymbol{y}$) by $\boldsymbol{V}_s$, the noncentrality τ parameter of a noncentral F distribution with degrees of freedom a and b by $H(a, b, \alpha, \tau)$. Assume $\mathbf{G}$: $(S-1) \times 1$ is group indicator vector whose sth element is 1 corresponding to treatment s; 0 otherwise. Denote $Cov(\hat{\boldsymbol{\beta}}) = C^{-1}\boldsymbol{\Gamma^{-1}}$. The focus now is on testing a set of linear hypotheses related to group-by-time (or a function of time) interaction parameters which are expressed in the following general linear hypothesis set up of the fixed-effect parameters $\boldsymbol{\beta}$, $H_0 : \boldsymbol{L}\boldsymbol{\beta} = 0 \; vs. \; H_1 : \boldsymbol{L}\boldsymbol{\beta} \neq 0$.

4.2. Results

Assume that a study wants to compare S treatments in C centers utilizing a longitudinal design of length T, and an allocation vector of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)^t$. Further assume that each center wants to use n subjects and randomization is performed at the center level, i.e. all subjects in a given center receive the same treatment assigned to that particular center. Let the proportion of dropouts in centers receiving the sth treatment be $\boldsymbol{\xi}_s = (\xi_{s,1}, \ldots, \xi_{s,T})^t$. In order to attain at least $(1-\tau)100\%$ power for the test specified in H_1 at an alternative value of $\boldsymbol{\beta} = \boldsymbol{\beta^*}$, the required number of subjects n per center should maintain the following constraint:

$$n \;\geq\; \min\{j : \hat{\lambda}(j) \geq H(S-1, C-S, \alpha, \tau)/C\}, \tag{3}$$

where $\hat{\lambda}(j) = (\boldsymbol{L}\boldsymbol{\beta}^*)^t(\boldsymbol{L}\hat{\boldsymbol{\Gamma}}^{-1}\boldsymbol{L}^t)^{-1}(\boldsymbol{L}\boldsymbol{\beta}^*)$. An arbitrary value of C cannot provide a valid solution of (3). Equation (3) provides a feasible solution only when $C \geq C^*$, where

$$C^* = H(S-1, C-S, \alpha, \tau)/(\boldsymbol{\beta}^{*t}\boldsymbol{L}^t[\boldsymbol{L}((\boldsymbol{U}^t\boldsymbol{\Delta}_\pi\boldsymbol{U}) \otimes \boldsymbol{\Sigma}_\gamma^{-1})^{-1}\boldsymbol{L}^t]^{-1}\boldsymbol{L}\boldsymbol{\beta}^*), \tag{4}$$

where, $\boldsymbol{\Delta}_\pi$ is a diagonal matrix with diagonal elements π_s, $\boldsymbol{U} = (\boldsymbol{u}_1^t, \cdots, \boldsymbol{u}_s^t)^t$ and $\boldsymbol{u}_s = (1 \quad \boldsymbol{G}_s^t)^t$ and $\otimes$ is the Kronecker product. Thus, C^* is the lower bound of C and is independent of n. The proof is mathematically intensive and lengthy, hence is not given here (see Amatya and Bhaumik (2018) for complete derivation). This result suggests that at least C^* clusters are necessary for a cluster randomized study to achieve the desired level of power $1-\tau$. As C increases (starting from C^*), the requirement for the number of subjects decreases, provided all other parameters remain fixed. An explicit expression of C^*

is given in Amatya and Bhaumik (2018). In order to evaluate the flexibility of this exciting result, we did some robustness studies via simulations (i) changing symmetric distributions (of random effects) to right skewed gamma distributions, (ii) relaxing the constraint of equal sample sample sizes of every center to 10% variations. The simulated power (under various parametric combinations, attrition rates, and model violations of types (i) and (ii)) was never less than 76% when it was fixed at 80%. Comprehensive results are reported in Amatya and Bhaumik (2018). A testing procedure with inflated Type I error rates will require fewer samples, but such a test will often show significance when the intervention effect is actually non-significant. On the other hand, a very conservative test will require more resources to attain the same target power (*e.g.*, 80%) compared to an exact test. Our proposed procedure avoids both scenarios. In order to demonstrate how fatal it can be in terms of power, when inappropriate methods are used for sample size determination we compared our proposed method with two existing methods by Murray (1998) and Heo, *et al.* (2013) designed for linear models. Results are presented in Table where for various values of between-cluster variation in slopes ($\sigma^2_{\gamma 22}$) we compute cluster size and corresponding power. Note that power for both the existing methods is substantially lower than what was targeted at 80%.

Table 3: Comparison of required number of clusters estimated from Murray (1998), Heo, *et al.* (2013), the proposed method, and the power attained in simulated evaluation

$\sigma^2_{\gamma 22}$	Murray (1998)		Heo, *et al.* (2013)		Proposed	
	C	power	C	power	C	power
0.03	8	.287	12	.367	35	.797
0.04	8	.252	12	.361	41	.797
0.05	8	.237	12	.339	46	.773
0.06	8	.243	12	.297	52	.778
0.07	8	.223	12	.275	58	.767
0.08	8	.218	12	.279	64	.795
0.35	10	.153	12	.216	224	.799

Both the existing methods perform well when outcome is linear, however, they are inappropriate for non-linear outcomes. Hence, sample size methodologies should be developed taking into account all complexities (type of outcome, within and between cluster variation, attrition rate) which is incorporated in our proposed method.

5. Conclusions

Health service utilization researchers regularly conduct multi-center studies which are longitudinal in nature. In these studies multiple correlated measurements are obtained from subjects who are nested within hospitals, schools etc. The distribution of the outcome variable usually is highly skewed with a profusion of zero as a large majority of eligible subjects never utilize service either due to lack of need or access, and a long right tail as some subjects are mass consumers of service. Sample size estimation methods used to design these hierarchical longitudinal studies with skewed zero-inflated outcome data either rely on completely inappropriate linear models or employ simple designs ignoring various levels of hierarchy which can result in severe under-estimation of resulting power. We derive a robust method for sample size estimation that incorporates multiple random-effects in a zero-inflated model. Our simulation study showed the proposed method achieved the desired 80% power consistently whereas the other competing approaches under estimated the power severely. During the data analysis phase researchers are routinely forced to exclude important random-effects from their fitted models due to model convergence issue. We propose a novel technique based upon Laplace approximation which considerably reduces the non-convergence and utilizes less computing resources in comparison to the existing methods.

References

Amatya, A. and Bhaumik, D. K. (2018). Sample size determination for multilevel hierarchical designs using generalized linear mixed models. *Biometrics*, **74**, 673–684.

Amatya, A., Bhaumik, D. K. and Gibbons, R. D. (2013). Sample size determination for clustered count data. *Statistics in Medicine*, **32**, 4162–4179.

Amatya, A., Bhaumik, D. K., Normand, S-L.T., Greenhouse, J., Kaizar, E., Neelon, B. and Gibbons, R. D. (2015). Likelihood-based random-effect meta-analysis of binary events. *Journal of Biopharmaceutical Statistics*, **25**, 984–1004.

Atkins, M. S., Shernoff, E. S., Frazier, S. L., Schoenwald, S. K., Cappella, E., Marinez-Lora, A., Mehta, T. G., Lakind, D., Cua, G., Bhaumik, R. and Bhaumik, D. K. (2015). Re-Designing community mental health services for urban children: supporting schooling to promote mental health. *Journal of Consulting Clinical Psychology*, **83**, 839–852.

Bandyopadhyay, D., DeSantis, S. M., Korte, J. and Brady, K. T. (2011). Some considerations for excess zeroes in substance abuse research. *American Journal of Drug and Alcohol Abuse*, **37**, 376–382.

Bhaumik, D. K., Roy, A., Aryal, S., Hur, K., Duan, N., Normand, S-L.T., Brown, C. H. and Gibbons, R. D. (2008). Sample size determination for studies with repeated continuous outcomes. *Psychiatric Annals*, **38**, 765– 771

Bhaumik, D. K., Amatya, A., Normand, S-L.T, Greenhouse, J., Kaizar, E., Neelon, B. and Gibbons, R. D. (2012). Meta- analysis of rare binary adverse event data. *Journal of American Statistical Association*, **107**, 555–567.

Bhaumik, D. K., Aryal, S., Amatya, A., Kapur, K. and Gibbons, R. D. (2011). Sample size determination for between group comparisons in mixed effects logistic regression models for analysis of longitudinal data. *Journal of Applied Statistical Science*, **19**, 11–22.

Bhaumik, D.K., Aryal, S. and Hur, S. (2019). How to select a suitable model for analysis of mental health service use data. *IAPQR Transactions*, Manuscript accepted for Publication.

Chen, P., Wu, T. J. and Yang, J. (2008). A comparative study of model selection criteria for the number of signals. *IET Radar, Sonar and Navigation*, **2**, 180–188.

Cook, J. A., Shore, S., Burke-Miller, J. K., Jonikas, J. A., Hamilton, M., Ruckdeschel, B., Norris, W., Markowitz, A. F., Ferrara, M. and Bhaumik, D. K. (2019). Mental health self-directed care financing: efficacy in improving outcomes and controlling costs for adults with serious mental illness. *Psychiatric Services*, **70**, 191–201.

Diehr, P., Yanez, D. and Ash, A., Hornbrook, M., and Lin, D. Y. (1999). Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, **20**, 125–144.

Donner, A. and Klar, N. (2002). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.

Ghosh, S. K., Mukhopadhyay, P. and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**, 1360–1375.

Gilbert, A. R., Domino, M. E. , Morrissey, J. P. and Gaynes, B. N. (2012). Differential service utilization associated with trauma-informed integrated treatment for women with co-occurring disorders. *Administration and Policy in Mental Health and Mental Health Services Research*, **39**, 426–439.

Gupta, R., Szczesniak, R. D. and Macaluso, M. (2015). Modeling repeated count measures with excess zeros in an epidemiological study. *Annals of Epidemilogy*, **25**, 583–589.

Hacker, K. A., Penfold, R. B., Arsenault, L.S., Zhang, F., Murphy, M. and Wissow, L. S. (2014). Behavioral health services following implementation of screening in massachusetts medicaid children. *Pediatrics*, **134**, 737–746.

Heo, M., Xue, X. and Kim, M. Y. (2013). Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials with random slopes. *Computational Statistics and Data Analysis*, **60**, 169–178.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York, NY.

Kapur, K., Bhaumik, R., Tang, X-C., Hur, K., Reda, D. J. and Bhaumik, D. K. (2014). Sample size determination for longitudinal designs with binary response. *Statistics in Medicine*, **33**, 3781–3800.

Kass, R. E. and Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association*, **90**, 773–795.

Lindsey, M. A., Brandt, N. E., Becker, K. D., Lee, B. R., Barth, R. P., Daleiden, E. L. and Chorpita, B. F. (2014). Identifying the common elements of treatment engagement interventions in children's mental health services. *Clinical Child and Family Pscyhology Review*, **17**, 283–298.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models* (2nd edition). Chapman and Hall, New York.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K. and McArdle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, **35**, 1193–1202.

Mullahy, J. (1986). Specification and test of ome modified count data models. *Journal of Econometrics*, **3**, 341–365.

Overall, J. E. and Doyle, S. R. (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials*, **15**, 100–123.

Pardoe, I. and Durham, C. A. (2003). Model choice applied to consumer preferences. *In Proceedings of the Joint Statistical Meetings*. American Statistical Association.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, **2**, 173–185.

Raudenbush, S. W. and Liu, X. F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, **5**, 199–213.

Ridout, M., Hinde, J. and Demetrio, C. G. (2001). A score test for testing a zero-in ated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219–223.

Roy, A., Bhaumik, D. K., Aryal, S. and Gibbons, R. D. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, **63**, 699–707.

Saha, A. and Dong, D. (1997). Estimating nested count data models. *Oxford Bulletin of Economics and Statistics*, **59**, 423–430.

Seghouane, A. K. (2006). Multivariate regression model selection from small samples using Kullback's symmetric divergence. *Signal Processing*, **86**, 2074–2084.

Tu, S. and Xu, L. (2012). A theoretical investigation of several model selection criteria for dimensionality reduction. *Pattern Recognition Letters*, **33**, 1117–1126.

Xie, H., Tao, J., McHugo, G. J. and Drake, R. E. (2013). Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. *Journal of Substance Abuse Treatment*, **45**, 99–108.