



Published in final edited form as:

Stat Med. 2014 June 15; 33(13): 2238–2250. doi:10.1002/sim.6091.

A hierarchical finite mixture model that accommodates zero-inflated counts, non-independence, and heterogeneity

Charity J. Morgan, PhD,

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294

Mark F. Lenzenweger, PhD,

Department of Psychology, State University of New York at Binghamton, Binghamton, New York 13902

Donald B. Rubin, PhD, and

Department of Statistics, Harvard University, Cambridge, MA 02138

Deborah L. Levy, PhD

Psychology Research Laboratory, McLean Hospital, 115 Mill Street, Belmont, MA 02478

Abstract

A number of mixture modeling approaches assume both normality and independent observations. However, these two assumptions are at odds with the reality of many data sets, which are often characterized by an abundance of zero-valued or highly skewed observations as well as observations from biologically related (i.e., non-independent) subjects. We present here a finite mixture model with a zero-inflated Poisson regression component that may be applied to both types of data. This flexible approach allows the use of covariates to model both the Poisson mean and rate of zero-inflation and can incorporate random effects to accommodate non-independent observations. We demonstrate the utility of this approach by applying these models to a candidate endophenotype for schizophrenia, but the same methods are applicable to other types of data characterized by zero inflation and non-independence.

Keywords

Zero-inflated Poisson regression; mixture modeling; psychopathology; endophenotype; thought disorder

1. Introduction

One of the major statistical analytic challenges faced by researchers concerns the distributions of observed count values characterized by a plethora of zero values (or a functional equivalent, lowest possible values). Examples include rating scale data (e.g., symptom severity) and objectively measured neurocognitive or motor processes where zero values occur frequently and are thought to reflect either the absence of pathology or poor

performance, respectively. This abundance of “zero” values typically results in markedly skewed positive distributions with long extended right tails. Clearly, zero-inflated data produce distributions that fall well short of normality. This type of data is not well-modeled using methods developed for normally distributed data [1, 2]; the application of such techniques to zero-inflated count data can yield biased, inefficient results [3]. Further, when multiple variables are characterized by actual or potential zero-inflation, multivariate normality cannot be expected to model these data well. Thus, count data with skewed distributions with an excess of zero values point to a need for data analytic methods that explicitly take the possibility of zero-inflation into account.

The application of finite mixture modeling to data from psychopathology investigations has been gaining momentum over the years. The utility of the mixture modeling approach, especially as contrasted with traditional clustering methods and other data reduction approaches, has been reviewed elsewhere [4]. Despite its utility and potential (c.f., Gibbons et al. [5]; Levy et al. [6]), mixture modeling has seen only modest application in psychopathology research [4] and usually focuses on only one index of interest (i.e., univariate mixture analysis). Examples include psychosis-proneness [7], ventricle size [8], age-at-admission [9], and smooth pursuit eye movement parameters [10]. Mixture modeling methods have also been developed to take into account both between-subject and within-subject factors that impact performance on laboratory tasks [11–14].

As interest in laboratory methods for the measurement of putative psychiatric endophenotypes [15] has grown, especially in connection with efforts to resolve the nature and influence of genetic factors in liability for illness, the utility of finite mixture modeling approaches has become more salient [16]. Family data are especially relevant to the analysis of putative endophenotypes, because first-degree biological relatives of affected individuals would be expected to be a mixture of gene carriers and non-gene carriers if the trait is subject to genetic influences. In order to capitalize on data from multiple individuals in the same family, mixture models that take into account the non-independence of observations from biologically related individuals are a necessity. Early mixture modeling approaches typically made two core assumptions: multivariate normality, and statistical independence (or conditional independence given covariates) of observations. These two assumptions are not met in many data sets, which are often zero-inflated and include observations from biologically related subjects, both of which affect the structure of the data. Thus, in order for many areas of research to take full advantage of finite mixture modeling, there is a need for methods that simultaneously take into account both zero-inflation in data values and the non-independence of observations within a sample.

In this paper, we present a series of mixture models that can be applied to zero-inflated count data. This method is useful not only for the specific example of count data with an over-abundance of zeros, but also for more general applications that result in data that, while not strictly being count data, can nevertheless be well-modeled using a zero-inflated Poisson (ZIP) distribution (e.g., values obtained from a symptom severity rating scale). In Section 2, we review the definition and notation for the zero-inflated Poisson distribution, a distribution that is well equipped to handle zero-inflated observations. In Section 3, we discuss finite mixture models and in Section 4 we present the ZIP mixture model, a model

that can be applied to zero-inflated data with non-independent observations. Section 5 contains details on how to fit and choose from among these models as well as information about how to assess the goodness of fit of a model. In Section 6 we apply these models to a candidate endophenotype for schizophrenia.

2. The Zero-inflated Poisson Distribution

In this paper we use the zero-inflated Poisson (ZIP) distribution to model event count data that may contain more zero-valued observations than would be expected for data arising from a Poisson distribution. The definition of the ZIP distribution is as follows:

Definition

If a random variable Y with probability π equals zero and with probability $(1-\pi)$ follows a Poisson distribution with mean λ , that is,

$$P(Y=0)=\pi+(1-\pi)e^{-\lambda}$$

$$P(Y=y)=(1-\pi)\frac{\lambda^y e^{-\lambda}}{y!}, \text{ for } y>0,$$

then we say that Y is distributed as a zero-inflated Poisson with Poisson mean λ and probability of zero-inflation π and use the following notation: $Y \sim \text{ZIP}(\lambda, \pi)$. Event count data that are distributed as ZIP can be conceptualized as arising from one of two sources: 1) a proportion π of the time that no event will occur; 2) for the remaining $1-\pi$ proportion of the time that the number of events arises from a Poisson process with mean λ (see also Ridout et al [17] for an overview of methods for zero-inflated count data).

Lambert [18] developed the ZIP regression model, which allows both the Poisson mean and probability of zero-inflation to depend on covariates. Lee et al. [19] extended this model to accommodate repeated measures and/or clustered data, adding the flexibility obtained by including random effects:

$$Y_{ij} \sim \text{ZIP}(\lambda_{ij}, \pi_{ij})$$

$$\log(\lambda_{ij}) = \beta \cdot x_{ij} + u_{ij}$$

$$\text{logit}(\pi_{ij}) = \psi \cdot x_{ij} + w_{ij}$$

$$u_{ij} \sim N(0, \sigma_u^2), w_{ij} \sim N(0, \sigma_w^2),$$

where, x_{ij} is a vector of covariates for the j^{th} subject within the i^{th} cluster. We note here that the same set of covariates is not necessarily used to model λ and π . Cluster-level random effects can easily be included in the model, for example by setting $u_{ij} = u_i$ for all subjects within the i^{th} cluster.

For other examples of hierarchical Poisson models, see Tsutakawa [20], Christiansen and Morris [21], or Geoffroy and Weerakkody [22].

3. Finite Mixture Models

For data with n observations, let Y_{ij} be the outcome for the j^{th} subject within the i^{th} cluster. Let the probability density function of Y_{ij} be

$$f(y_{ij}) = \sum_{k=1}^m p_{ij,k} \cdot g_k(y_{ij}),$$

$$\sum_{k=1}^m p_{ij,k} = 1 \text{ and } p_{ij,k} \geq 0, \forall k,$$

where m is the number of components or risk classes, $p_{ij} = (p_{ij,1}, \dots, p_{ij,m})$ is the vector of mixing proportions, $p_{ij,k}$ is the mixing proportion for the j^{th} subject within the i^{th} cluster and component k , and the g_k are density functions [23–25]. When Y_i follows this distribution, we can interpret the data Y as arising from a mixture of m components, or risk classes, where the density of the k^{th} component is $g_k(\cdot)$. For this paper we turn our attention to the specific case where the $g_k(\cdot)$ are either Poisson or ZIP density functions.

Although for some applications, it may be appropriate to assume that the prior probability of belonging to a given risk class is the same for all individuals, the vector of mixing proportions, p_{ij} , need not be identical for all subjects. For example, a first-degree relative of an individual with a disease may be *more* likely to show abnormal performance or an abnormal trait than a person with no family history of that disease; indeed, such a difference is one of the criteria for an endophenotype. For such applications, the mixing proportions can be allowed to depend on covariates, typically by modeling p_{ij} using multinomial logistic regression:

$$\log \left(\frac{p_{ij,k}}{p_{ij,1}} \right) = \gamma_k x_{ij}$$

$$k=2, \dots, m$$

[25, 26].

Note that while it is not required that the same covariates be used to model the Poisson mean, rate of zero-inflation, and the mixing proportions, there may exist applications where it is necessary to include a subset of the covariates in all three regression models. The methods presented here are still applicable in these settings.

In the finite mixture models described above we assumed that the number of components, m , is a known quantity. Although methods for estimating the number of components do exist [27–29] [27–29], in this paper we take the more common approach of fitting multiple finite mixture models with varying values of m and then comparing the fits of those models. See also MacLachlan and Khan [30] for a comparison of methods for selecting the number of components.

4. The ZIP Mixture Model

In this section we describe a model that can be used to fit heterogeneous, zero-inflated count data. This model takes into account two possible sources of heterogeneity: 1) the heterogeneity resulting from the presence of distinct subpopulations, or mixture components, and 2) the heterogeneity arising from variability within those subpopulations. We model the data using a finite mixture model composed of m classes. Without loss of generality, we order the classes so that the probability, or “risk,” of an event increases with the class label. That is, subjects belonging to the first risk class are at lowest risk and subjects belonging to the m^{th} component have the highest probability of an event. In order for the model to be statistically identifiable (i.e., parameters for the model are estimable), only one component can be subject to zero-inflation. Since zero-inflation results in an event not occurring, it is reasonable to assume that those subjects who are susceptible to zero-inflation should be at low risk for the event (assuming that zero values reflect the most normal score). Thus, we assume that only subjects belonging to the first class are subject to zero-inflation; observations from these subjects are modeled using a ZIP regression.¹ Observations arising from each of the remaining risk classes are assumed to follow Poisson distributions with increasing means.

A random effects structure such as the one suggested by Lee et al [19] is incorporated into all Poisson and ZIP regression models to handle the presence of non-independent observations in the data (e.g., repeated measurements taken on the same individual or data obtained from members of the same family).²

The structure of this model is summarized in Table 1:

The ZIP mixture model belongs to the larger class of mixture regression models (see Wedel and DeSarbo [31] for a review) and is an extension of a model proposed by Lenk and DeSarbo [32], who noted that an approach that combines finite mixture modeling with mixed effect regression modeling could well model data comprised of distinct, heterogeneous subpopulations or mixture classes.

5. Model Fitting and Comparison

In taking a Bayesian approach, we use the posterior distributions of the model parameters to make inferences. Guidance on Bayesian methods for finite mixture models can be found in Lenk and DeSarbo [32]. Models are compared using the Bayesian Information Criterion (BIC; [33]). See Nagin [34] for a discussion of the use of BIC to select the number of components for a finite mixture model. When comparing models using the BIC, the model that yields the smallest BIC value when fitted to the data is selected as the best-fitting.³

¹There may exist applications where individuals at high risk are subject to zero-inflation. The methods presented here can easily be adapted to fit such situations.

²The random effects model can easily be adapted to a wide variety of situations via the inclusion of covariates. For example, an autoregressive model may be used to model the temporal dependence of repeated measurements.

³When models are nested, they may also be compared using the Likelihood Ratio Test (LRT). However, standard asymptotic p -values have been shown not to be appropriate when using the LRT to compare finite mixture models [35, 36]. For these cases, posterior predictive checks [37] can be used to create a reference distribution for the test statistic for the LRT (see also Lo, Mendell, and Rubin [38]).

Once a final model has been selected, the goodness of fit of that model can be assessed using posterior predictive checks (PPC) [37] to determine whether the model reproduces key features of the data. Under this approach, a test statistic that captures some important aspect of the data is selected. Parameter values are drawn from their posterior distribution and a replicate of the data is simulated using the drawn values. The chosen test statistic is then calculated from the replicated data. The observed value of the test statistic is compared to a sample of test statistics calculated from replicated data sets. For each statistic, a posterior predictive p -value is calculated by computing the proportion of simulated statistics that are at least as extreme as the observed value. A well-fitting model will yield several replicated statistics that are comparable to the observed value, and thus the associated p -value will be large.

6. Example

We demonstrate this approach by applying these models to an example relevant to schizophrenia family data. Linkage and association studies of schizophrenia have so far yielded weak and inconsistent results [39–41]. Indeed, common variants account for only about a third of the genetic variance in risk for schizophrenia and show substantial shared genetic liability for bipolar disorder [42]. Such pleiotropic effects of specific SNPs were recently confirmed across both adult and childhood onset psychiatric disorders [43]. A number of traits have been identified that are both associated with schizophrenia and that aggregate in relatives of schizophrenia patients at a rate much higher than that of the clinical disorder. These traits, provisionally identified *endophenotypes*, may be alternative manifestations of schizophrenia risk genes that are more penetrant than schizophrenia itself. These endophenotypes may help to identify relatives who are non-penetrant carriers of one or more of the genes that increase susceptibility for schizophrenia, thereby improving power to detect disease loci. In this worked example, we use thought disorder as a candidate endophenotype for schizophrenia.

Thought disorder was described by Bleuler [44] as a “loosening of associations” and both Bleuler [44] and Kraepelin [45] considered this symptom to be a core feature of schizophrenia psychopathology. An individual who suffers from thought disorder may exhibit difficulties in concept formation, cognitive focus, reasoning, and/or reality testing [46]. Meehl [47, 48] suggested that mild thought disorder should also be found in those who harbor the genetic liability for schizophrenia, but who may never display the clinical disorder (see also Lenzenweger [16]; Levy et al. [49]).

The data set considered in these analyses included 286 participants. Of these participants, 173 (66 males and 107 females) were first-degree biological relatives of schizophrenia and schizo-affective patients (RelSZ) who did not meet diagnostic criteria for a psychotic disorder, bipolar disorder without psychotic features, or a schizophrenia spectrum personality disorder (schizoid, schizotypal, or paranoid). A detailed description of the clinical assessment and diagnostic procedures can be found elsewhere [50, 51]. The non-psychiatric control (NC) group was comprised of 113 individuals (45 males and 68 females). The NC met the same clinical exclusion criteria described for RelSZ above but also did not have a family history of psychosis, psychiatric hospitalization, or suicide. Demographic

characteristics of the subject groups are presented in Table 1. The two groups did not significantly differ in estimated verbal IQ ($p = 0.29$), socio-economic status (SES, $p = 0.44$), years of education ($p = 0.75$), or sex ratio ($p = 0.87$). The NC group was significantly younger ($M = 42.78$ years) than the RelSZ group ($M = 49.27$, $p < 0.001$).

The majority of the participants (63.41%) were related to at least one other participant in the study. The RelSZ group consisted of a total of 88 families. Three families had five members each, four families had four members each, 19 families had three members each, 23 families had two members each and 39 families had one member. In the NC group there were a total of 84 families. Four families had four members each, two families had three members each, 13 families had two members each, and 65 families consisted of one member.

Measures

Verbal responses to ten cards of the Rorschach Inkblot [54] were tape recorded and transcribed verbatim. Three experienced raters scored the verbatim transcript for thought disorder using the Thought Disorder Index (TDI) [55, 56] without knowledge of group membership. The TDI is a reliable and valid scale used to assess twenty-three categories of thought disorder; the intra-class correlation coefficient (ICC) for four teams of raters for Total TDI score was 0.74; Spearman correlations among six pairs of rating teams were between 0.81 and 0.90 [55, 57]. Use of the TDI to characterize thought disorder in different patient groups and in relatives has been described extensively [55, 58–65] (see Holzman et al. [66] for a review). Four of these categories (peculiar verbalization, queer response, absurd response, neologism) refer to successively more severe idiosyncratic use of language, or “deviant verbalizations.” Deviant verbalizations are characteristic of the thought disorder associated with schizophrenia in all clinical states [58–60] and are also over-represented in clinically unaffected relatives [55, 59, 61, 63]. More deviant verbalizations signify stronger evidence of schizophrenia-related thought disorder. The number of deviant verbalizations in each subject’s verbatim responses was the outcome variable.

The RelSZ made a significantly higher number of deviant verbalizations ($\bar{x} = 4.0$, s.d. = 4.84) than the NC ($\bar{x} = 2.11$, s.d. = 3.84) ($p < 0.001$, Wilcoxon rank sum test). The two groups did not differ significantly in variance of deviant verbalizations scores ($p > 0.10$, $Z = 1.20$ from Miller’s [67] jackknife test for equality of variance). Figure 1 displays histograms showing the distributions of the deviant verbalization score in RelSZ and NC. At least one deviant verbalization was present in 128 (74.0%) of the RelSZ and in 60 NC (53.1%), a statistically significant difference ($\chi_{(1)}^2 = 12.33$, $p < 0.001$). Age was not significantly correlated with number of deviant verbalizations in either the RelSZ ($r = -0.08$, $p > 0.10$) or the NC ($r = 0.07$, $p > 0.10$). Among NC, male subjects ($\bar{x} = 3.44$, s.d. = 4.79) made significantly more deviant verbalizations than female subjects ($\bar{x} = 1.22$, s.d. = 2.75) ($p < 0.001$, Wilcoxon rank sum test); this relationship was not observed among RelSZ ($p > 0.10$). The distribution of deviant verbalizations in the RelSZ appears to be bimodal, suggesting that a finite mixture model may be suitable for these data. Figure 1 also shows the relatively large proportion of zero-valued observations (i.e., no deviant verbalizations) in both groups.

Method

Since this particular data set does not include any repeated measurements, we do not need to include any subject level effects into our model; however, because this data set contains related individuals, we incorporated cluster (i.e., family) level effects, u_i , in the model. We note that since the data consist of only first-degree relatives, a simple random effect for each family should be sufficient to handle the non-independence introduced by the inclusion of related individuals. Although not needed here, covariates could potentially be included into the random effects portion of the model to handle data comprised of subjects with differing degrees of relatedness (for example, second- versus first-degree relatives or dizygotic versus monozygotic twins).

Since we observed a sex difference in rate of deviant verbalizations in NC, and sex effects have been found in previous studies of thought disorder [68], for our model we allow the deviant verbalization rate to depend on sex,

$$\begin{aligned} \log(\lambda_{ijk}) &= \beta_k + \beta_s \cdot S_{ij} + u_i \\ u_i &\sim N(0, \sigma_u^2) \end{aligned}$$

where S is an indicator variable for sex.

Recall that we assume that only subjects belonging to the lowest-risk class are subject to zero-inflation. However, it may be the case that subjects with a high verbal IQ may be more likely to show no deviant verbalizations than those with lower verbal IQ. In order to test this hypothesis, we include a measure of verbal IQ in our model for the rate of zero-inflation:

$$\text{logit}(\pi_{ij}) = \psi_0 + \psi_v \cdot V_{ij}$$

where V_{ij} is the verbal IQ for the j^{th} subject in the i^{th} cluster.

We also note that equality constraints over groups on the mixing proportions would clearly not be an appropriate if the quantitative trait in question were, indeed, a valid schizophrenia endophenotype. For example, relatives may be a mixture of gene carriers and non-gene carriers (e.g., 60% and 40%, respectively) and controls may all be non-gene carriers. Or, controls may also be a mixture but with very different mixing proportions from those observed in RelSZ (e.g., 5% and 95%, respectively). We use the following multinomial logistic model for the mixing proportions:

$$\log\left(\frac{p_{ij,k}}{p_{ij,1}}\right) = \gamma_{k0} + \gamma_{k1} \cdot x_{ij} \quad k=2, \dots, m$$

where

$$x_{ij} = \begin{cases} 1 & \text{if } j^{\text{th}} \text{ subject within } i^{\text{th}} \text{ cluster is RelSZ} \\ 0 & \text{otherwise} \end{cases}$$

We can then examine the posterior distribution of the covariates γ_{kl} to determine whether different mixing proportions for the RelSZ and NC are necessary to adequately model these data.

We also fit Lambert's ZIP regression model [18] to these data:

$$\begin{aligned} Y_{ij} &\sim \text{ZIP}(\lambda_{ij}, \pi_{ij}) \\ \log(\lambda_{ij}) &= \beta_1 + \beta_s \cdot S_{ij} + \beta_x x_{ij} + u_i \\ \text{logit}(\pi_{ij}) &= \psi_0 + \psi_V \cdot V_{ij} \\ u_i &\sim N(0, \sigma_u^2) \end{aligned}$$

as well as a standard Poisson regression model that was identical to the ZIP regression model described above, except that it did not allow for zero inflation (i.e., $\pi_{ij} = 0$, for all i, j). Note that for these regression models, mixture within the RelSZ is not modeled. That is, for a given sex, all RelSZ are assumed to have the same Poisson mean. This assumption is equivalent to assuming that all RelSZ belong to the same risk class and is clearly inappropriate for an endophenotype since at most, half of relatives would be expected to be gene carriers.⁴ Thus, failure of the data to support the ZIP mixture model in favor of either the ZIP or Poisson regression model would be inconsistent with the idea that the level of risk varies among RelSZ according to genetic liability and would provide evidence against the use of deviant verbalizations as a schizophrenia endophenotype.

Results

The data for this example were analyzed using the openBugs software package [69]; an example of BUGS code useful for fitting these types of models is provided in the Appendix. The prior distributions for all model parameters were uniform over a suitably large interval. This approach was taken to ensure that all prior distributions were proper, while providing little prior information. More information on the specific priors used can be found in the Appendix. Convergence was assessed using the Gelman-Rubin statistic, \hat{R} [36]. The Gibbs sampler was used to simulate draws from the posterior distributions of the different models. Starting points for the Gibbs samplers were randomly selected and each chain was iterated until approximate convergence ($\hat{R} < 1.1$). We fit versions of the ZIP mixture model with different numbers of mixing components. The log posterior density, log-likelihood, and BIC at the posterior modes for each of these models are displayed in Table 3.

The ZIP mixture model with two components returned a smaller BIC (1329.15) than the one-component model (BIC = 1737.02), providing evidence that the data were better explained using a finite mixture model than a one-component model. The BIC for the model with two classes was also smaller than that of the model with three classes (BIC = 1343.46); thus we conclude that two mixture components were sufficient to adequately model the data and did not fit the ZIP mixture model with four or more risk classes.

⁴On average, at most half of first-degree relatives would be expected to be gene carriers if the endophenotype or disease is transmitted by a dominant gene. For other non-X-linked modes of transmission, the average proportion of gene carriers would be lower.

Table 3 also gives the results from fitting the ZIP and Poisson regression models. The BIC values from fitting these models were larger than that of the two-component model. We see that the ZIP mixture model provided a better fit to the data than both the ZIP regression model (BIC = 1698.34) and the Poisson regression model (BIC = 1999.22), further confirming the need for using a finite mixture model.

We now report results for the two-component ZIP mixture model. One of the two mixture components corresponded to a low-risk group (Risk Class 1) whose deviant verbalization counts follow a zero-inflated Poisson distribution; the other mixture component represented the high-risk group (Risk Class 2), whose observations were distributed as Poisson. We drew 5000 values of the model parameters from their posterior distribution; Table 4 gives the 95% posterior intervals for these parameters.

The estimated odds ratio for RelSZ versus NC for membership in the high-risk component is 3.42 (95% posterior interval: 1.72 to 7.80); as the posterior interval (PI) for the odds ratio did not contain one, we concluded that the relationship between family type (i.e., RelSZ vs. NC) and deviant verbalization risk is statistically significant. Participants who are at high risk for deviant verbalizations represent an estimated 27.6% of the first-degree RelSZ, and 10.1% of the NC. That RelSZ are indeed at higher risk than NC for deviant verbalizations provides support for the use of thought disorder as a schizophrenia endophenotype.

The effect of sex on the deviant verbalization rate can be estimated using the posterior distribution for β_S . We estimated that male subjects make on average 1.34 times as many deviant verbalizations as female subjects (95% PI for this effect: 1.08 to 1.68). The effect of verbal IQ on the rate of zero-inflation was not statistically significant (95% PI for ψ_V : -0.31 to 0.47). The estimated mean number of deviant verbalizations for a high-risk participant is 12.38 for males and 9.22 for females. The number of deviant verbalizations for a low-risk participant is assumed to follow a zero-inflated Poisson distribution with an estimated Poisson mean of 2.09 for males and 1.56 for females, and estimated rate of zero-inflation of 28.3%. 95% posterior intervals for the model parameters are given in Table 4.

We next performed PPC to assess the fit of the final model; we selected the maximum number of deviant verbalizations, the variance, and the percentage of zeros as the test statistics. We also evaluated the final model's ability to replicate the familial structure present in these data by performing PPC on two additional test statistics: the intra-family correlation coefficient and the median intra-family variance. Each of these statistics was calculated for the entire sample as well as for the RelSZ and NC separately. All of the PPC p-values are greater than 0.05, indicating that the final model can adequately reproduce all of the summary statistics considered. Thus, we conclude that the final model performs well for both family and global statistics.

7. Discussion

Datasets that contain a large proportion of zero-valued observations are common in many areas of research. Such data can be too skewed for statistical methods that assume normality to perform well, motivating the use of data analytic methods that accommodate zero-inflated data. Furthermore, datasets that include related family members will be characterized by

non-independent observations. Finite mixture models are a useful tool for resolving the heterogeneity in these types of data. Mixture analytic methods for including related individuals and families of different size have been developed [70, 71] based on models that are similar in structure to models developed for repeated-measures data [11, 13]. At the time the work described here was being done (2010), there appeared to be no commercial software available to fit correlated zero-inflated Poisson models using a Bayesian approach. However, it appears that Mplus has recently been expanded to include that functionality [72].

The models we present can be adapted to many different kinds of data sets by adjusting the number of risk classes, selecting which risk class is subject to zero-inflation, or by removing the zero-inflated component altogether. In addition, the need for each adjustment to the basic model can be assessed by conducting a likelihood ratio test and using posterior predictive checks to test the resulting increase in likelihood for statistical significance. That these models can be further adjusted to allow for the analysis of correlated data provides additional flexibility and has considerable utility.

This method offers an advantage over the classic Poisson or ZIP regression model in that heterogeneity within groups is explicitly modeled. This feature is especially helpful when analyzing data from a proposed endophenotype. Use of the ZIP mixture model allows us to estimate the proportion of subjects who exhibit an abnormal behavior or trait, and thus, may be potential gene carriers.

The method presented here should only be applied to data that can be well modeled by a mixture of Poisson models with a zero-inflated component. Furthermore, in order for the parameters of the model to be estimable, only one of the mixture components may be zero-inflated; thus this model should not be applied to data that are subject to multiple sources of zero-inflation. Several steps can be taken to avoid the inappropriate application of the ZIP mixture model. Much work has been devoted to verifying the appropriateness of the Poisson assumption [73–75] as well as correctly identifying zero-inflation [19, 76]. We also advocate first fitting a version of the model with only one component and comparing the fit of that model to versions with two components to determine whether a finite mixture model is necessary to adequately model the data. Finally, as demonstrated above, posterior predictive checks [37] can be useful in assessing model fit and diagnosing model misspecifications. This approach can be applied to any dataset that meets the above criteria for a ZIP mixture model and can incorporate random effects and covariates as needed.

In our worked example, we apply our method to a provisionally identified endophenotype – a quantitative measure of “thought disorder with schizophrenic features” – in samples of clinically unaffected first-degree relatives of schizophrenia patients and nonpsychiatric controls. In the finite mixture modeling of the thought disorder data, we started with relatively simple models and successively incorporated more complex models in order to illustrate the model testing sequence and strategy and how to evaluate the goodness of fit of the various models to the data. This approach showed that a mixture model succeeds in fitting the deviant verbalizations data and that simple models do not fit the data as well as more complex ones. We also demonstrated that the probability of being identified at high

risk for thought disorder, based on the best-fitting model, is significantly greater for first-degree relatives of schizophrenics than for controls. These results provide support for the potential usefulness of deviant verbalizations as a schizophrenia endophenotype.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Michael J. Coleman and the late Dr. Philip S. Holzman for scoring thought disorder protocols, Olga Krastoshevsky for database assistance, and Michael J. Coleman, Donghyung Lee, Steven Matthyse, and Nancy R. Mendell for helpful comments on the paper.

Funding

This work was supported by the National Institute of Mental Health (R01 MH49487, MH071523 and MH31340); the Sidney R. Baer, Jr. Foundation; the Essel Foundation, and the National Association for Research on Schizophrenia and Depression (Brain and Behavior Research Foundation).

References

1. Cameron, AC.; Trivedi, PK. Regression Analysis of Count Data. 6. Cambridge University Press; New York: 2007.
2. Winkelmann, R. Econometric Analysis of Count Data. 5. Springer-Verlag; Berlin: 2008.
3. Atkins DC, Gallop RJ. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology*. 2007; 21:726–735. [PubMed: 18179344]
4. Lenzenweger MF, McLachlan G, Rubin DB. Resolving the latent structure of schizophrenia endophenotypes using expectation-maximization-based finite mixture modeling. *Journal of Abnormal Psychology*. 2007; 116:16–29. [PubMed: 17324013]
5. Gibbons RD, Dorus E, Ostrow D, Pandey GN, Davis JM, Levy DL. Mixture distributions in psychiatric research. *Biological Psychiatry*. 1984; 19:935–961. [PubMed: 6477998]
6. Levy DL, Holzman PS, Matthyse S, Mendell NR. Eye tracking dysfunction and schizophrenia: A critical perspective. *Schizophrenia Bulletin*. 1993; 19:461–536. [PubMed: 8235455]
7. Lenzenweger MF, Moldin SO. Discerning the latent structure of hypothetical psychosis proneness through admixture analysis. *Psychiatry Research*. 1990; 33:243–257. [PubMed: 2243900]
8. Daniel DG, Goldberg TE, Gibbons RD, Weinberger DR. Lack of a bimodal distribution of ventricular size in schizophrenia: A Gaussian mixture analysis of 1056 cases and controls. *Biological Psychiatry*. 1991; 30:887–903. [PubMed: 1747437]
9. Welham J, McLachlan G, Davies G, McGrath J. Heterogeneity in schizophrenia; mixture modelling of age-at-first-admission, gender and diagnosis. *Acta Psychiatrica Scandinavica*. 2000; 101:312–317.
10. Ross RG, Olincy A, Mikulich SK, Radant AD, Harris JG, Waldo M, Compagnon N, Heinlein S, Leonard S, Zerbe G, Adler L, Freedman R. Admixture analysis of smooth pursuit eye movements in probands with schizophrenia and their relatives suggests gain and leading saccades are potential endophenotypes. *Psychophysiology*. 2002; 39:809–819. [PubMed: 12462508]
11. Belin TR, Rubin DB. The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*. 1995; 14:747–768. [PubMed: 7644856]
12. Matthyse S, Levy DL, Wu Y, Rubin DB, Holzman PS. Modeling intermittent degradation in schizophrenic performance. *Schizophrenia Research*. 1999; 40:131–146. [PubMed: 10593453]
13. Rubin DB, Wu YN. Modeling schizophrenic behavior using general mixture components. *Biometrics*. 1997; 53:243–261. [PubMed: 9147594]

14. Lenzenweger M, Jensen S, Rubin D. Finding the “genuine” schizotyp: A model and method for resolving heterogeneity in performance on laboratory measures in experimental psychopathology research. *Journal of Abnormal Psychology*. 2003; 112:457–468. [PubMed: 12943024]
15. Gottesman II, Gould TD. The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry*. 2003; 130:636–645. [PubMed: 12668349]
16. Lenzenweger, MF. *Schizotypy and Schizophrenia: The View from Experimental Psychopathology*. Guilford; New York: 2010.
17. Ridout M, Demétrio CG, Hinde J. Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference*. 1998:179–192.
18. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34:1–14.
19. Lee AH, Wang K, Scott JA, Yau KKW, McLachlan GJ. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*. 2006; 15:47–61. [PubMed: 16477948]
20. Tsutakawa RK. Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*. 1988; 83:37–42. [PubMed: 12155410]
21. Christiansen CL, Morris CN. Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*. 1997; 92:618–632.
22. Geoffroy P, Weerakkody G. A Poisson-gamma model for two-stage cluster sampling data. *Journal of Statistical Computation and Simulation*. 2001; 68:161–172.
23. Titterington, DM.; Smith, AFM.; Makov, UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley; New York: 1985.
24. Everitt, BS.; Hand, DJ. *Finite Mixture Distributions*. Chapman and Hall; London: 1981.
25. McLachlan, GJ.; Peel, D. *Finite Mixture Models*. Wiley; New York: 2000.
26. Dayton CM, Macready GB. Concomitant-variable latent-class models. *Journal of the American Statistical Association*. 1988; 83:173–178.
27. McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*. 1987; 36:318–324.
28. Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 1997; 59:731–792.
29. Stephens M. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Annals of Statistics*. 2000; 28:40–74.
30. McLachlan GJ, Khan N. On a resampling approach for tests on the number of clusters with mixture model based clustering of tissue samples. *Journal of Multivariate Analysis*. 2004; 90:90–105.
31. Wedel, M.; DeSarbo, WS. A review of recent developments in latent structure regression models. In: Bagozzi, RP., editor. *Advanced Methods of Marketing Research*. Blackwell Publishing; Cambridge, MA: 1994.
32. Lenk PJ, DeSarbo WS. Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*. 2000; 65:93–119.
33. Schwarz GE. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
34. Nagin DS. Analyzing developmental trajectories: A semiparametric group-based approach. *Psychological Methods*. 1999; 4:139–157.
35. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*. 1996; 6:733–807.
36. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*. 1992; 7:457–511.
37. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*. 1984; 12:1151–1172.
38. Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika*. 2001; 88:767–778.
39. Risch N. Genetic linkage and complex diseases, with special reference to psychiatric disorders. *Genetic Epidemiology*. 1990; 7:3–16. [PubMed: 2184091]

40. Matthyse S. Genetic linkage and complex diseases: A comment. *Genetic Epidemiology*. 1990; 7:29–31.
41. Goldstein DB. Common genetic variation and human traits. *New England Journal of Medicine*. 2009; 360:1696–1698. [PubMed: 19369660]
42. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009; 460:748–751. [PubMed: 19571811]
43. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*. 2013; 381:1371–1379. [PubMed: 23453885]
44. Bleuler, E. *Dementia Praecox or the Group of Schizophrenias*. International Universities Press; New York: 1911/1950.
45. Kraepelin, E. *Dementia Praecox and Paraphrenia*. Chicago Medical Book Company; Chicago: 1896/1919.
46. Weiner, IB. *Psychodiagnosis in Schizophrenia*. Wiley; New York: 1966.
47. Meehl PE. Schizotaxia, schizotypy, schizophrenia. *American Psychologist*. 1962; 17:827–838.
48. Meehl PE. Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *Journal of Personality Disorders*. 1990; 4:1–99.
49. Levy DL, Coleman MJ, Sung H, Ji F, Matthyse S, Mendell NR, Titone D. The genetic basis of thought disorder and language and communication disturbances in schizophrenia. *Journal of Neurolinguistics*. 2010; 23:176–192. [PubMed: 20161689]
50. Brownstein J, Krastoshevsky O, McCollum C, Kundamal S, Matthyse S, Holzman PS, Mendell NR, Levy DL. Antisaccade performance is abnormal in schizophrenia patients but not in their biological relatives. *Schizophrenia Research*. 2003; 63:13–25. [PubMed: 12892854]
51. Coleman MJ, Titone D, Krastoshevsky O, Krause V, Huang Z, Mendell NR, Eichenbaum HE, Levy DL. Reinforcement ambiguity and novelty do not account for transitive inference deficits in schizophrenia. *Schizophrenia Bulletin*. 2010; 36:1187–1200. [PubMed: 19460878]
52. Wechsler, D. *Manual for the Adult Intelligence Scale-Revised*. Psychological Corporation; New York: 1981.
53. Hollingshead, AB. *Two factor index of social position*. Yale University Press; New Haven, CT: 1965.
54. Rorschach, H. *Psychodiagnostics*. Grune & Stratton; New York: 1921/1942.
55. Johnston, MH.; Holzman, PS. *Assessing Schizophrenic Thinking*. Jossey-Bass, Inc; San Francisco: 1979.
56. Solovay M, Shenton M, Gasperetti C, Coleman M, Kestnbaum E, Carpenter J, Holzman P. Scoring manual for the thought disorder index. *Schizophrenia Bulletin*. 1986; 12:483–496. [PubMed: 3764364]
57. Coleman MJ, Carpenter T, Waternaux C, Levy DL, Shenton M, Perry J, Medoff D, Wong H, Manoach D, Meyer P, O'Brian C, Valentino C, Robinson D, Smith M, Makowski D, Holzman PS. The thought disorder index: A reliability study. *Psychological Assessment*. 1993; 5:336–342.
58. Solovay M, Shenton M, Holzman P. Comparative studies of thought disorders. I. Mania and schizophrenia. *Archives of General Psychiatry*. 1987; 44:13–20. [PubMed: 3800579]
59. Holzman PS, Shenton M, Solovay M. Quality of thought disorder in differential diagnosis. *Schizophrenia Bulletin*. 1986; 12:360–372. [PubMed: 3764357]
60. Spohn HE, Coyne L, Larson J, Mittleman F, Spray J, Hayes K. Episodic and residual thought pathology in chronic schizophrenia. *Schizophrenia Bulletin*. 1986; 12:394–407. [PubMed: 2876514]
61. Kinney DK, Holzman PS, Jacobsen B, Jansson L, Faber B, Hildebrand W, Kasell E, Zimbalist ME. Thought disorder in schizophrenic and control adoptees and their relatives. *Archives of General Psychiatry*. 1997; 54:475–479. [PubMed: 9152101]
62. Shenton ME, Solovay MR, Holzman PS. Comparative studies of thought disorders. II. Schizoaffective disorder. *Archives of General Psychiatry*. 1987; 44:21–30. [PubMed: 3800580]
63. Shenton ME, Solovay MR, Holzman PS, Coleman MJ, Gale HJ. Thought disorder in the relatives of psychotic patients. *Archives of General Psychiatry*. 1989; 46:897–901. [PubMed: 2489936]

64. Makowski DG, Wateraux C, Lajonchere CM, Dicker R, Smoke N, Koplewicz H, Min D, Mendell NR, Levy DL. Thought disorder in adolescent-onset schizophrenia. *Schizophrenia Research*. 1997; 23:147–165. [PubMed: 9061811]
65. Wahlberg K-E, Wynne LC, Oja H, Keskitalo P, Anais-Tanner H, Koistinen P, Tarvainen T, Hakko H, Lahti I, Moring J, Naarala M, Sorri A, Tienari P. Thought Disorder Index of Finnish adoptees and communication deviance of their adoptive parents. *Psychological Medicine*. 2000; 30:127–136. [PubMed: 10722183]
66. Holzman, PS.; Levy, DL.; Johnston, MH. The use of the Rorschach technique for assessing formal thought disorder. In: Bornstein, RF.; Masling, JM., editors. *Scoring the Rorschach: Seven Validated Systems*. Lawrence Erlbaum Associates; Mahwah, NJ: 2005.
67. Miller RG Jr. Jackknifing variances. *Annals of Mathematical Statistics*. 1968; 39:567–582.
68. Coleman MJ, Levy DL, Lenzenweger MF, Holzman PS. Thought disorder, perceptual aberrations and schizotypy. *Journal of Abnormal Psychology*. 1996; 105:469–473. [PubMed: 8772019]
69. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*. 2009; 28:3049–3067. [PubMed: 19630097]
70. Hasstedt, SJ. Pedigree analysis package. Department of Human Genetics, University of Utah; Salt Lake City, Utah: 1994.
71. Sorant, AJM.; Bonney, GE.; Elston, RC. REGC User's Manual. *Statistical Analysis for Genetic Epidemiology*. Vol. 2.1. Department of Biometry and Genetics, Louisiana State University Medical Center; New Orleans: 1992.
72. Muthén, LK.; Muthén, BO. Mplus User's Guide. Seventh. Muthén & Muthén; Los Angeles, CA: 1998–2012.
73. Collings BJ, Margolin BH. Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*. 1985; 80:411–418.
74. Dean CB, Lawless JF. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*. 1989; 84:467–472.
75. Xiang L, Lee AH, Yau KKW, McLachlan GJ. A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statistics in Medicine*. 2007; 26:1608–1622. [PubMed: 16794991]
76. van den Broek J. A score test for zero inflation in a Poisson distribution. *Biometrics*. 1995; 51:738–743. [PubMed: 7662855]

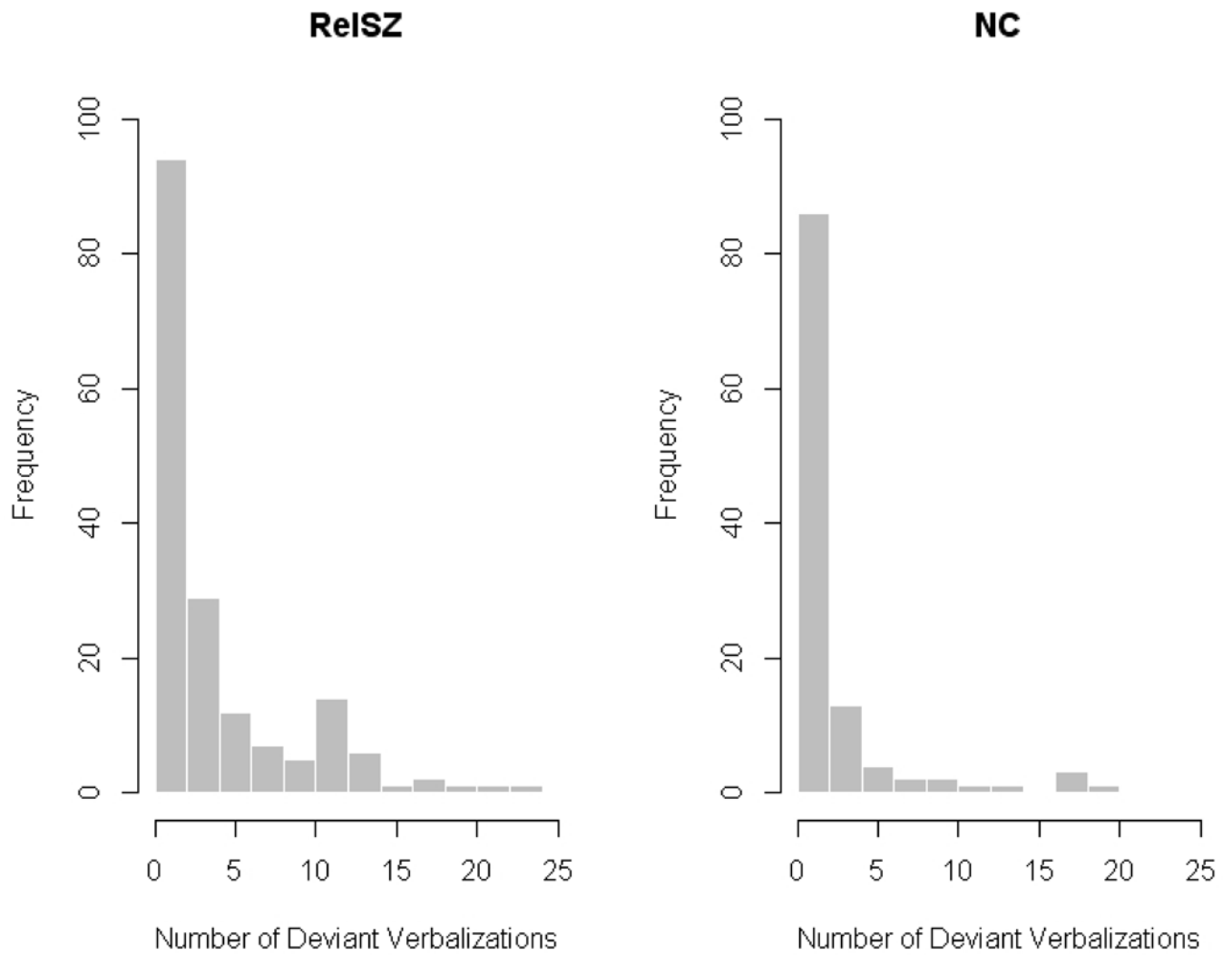


Figure 1. Histograms Comparing the Distribution of the Number of Deviant Verbalizations in First-Degree Relatives of Schizophrenia Patients and Non-psychiatric Controls

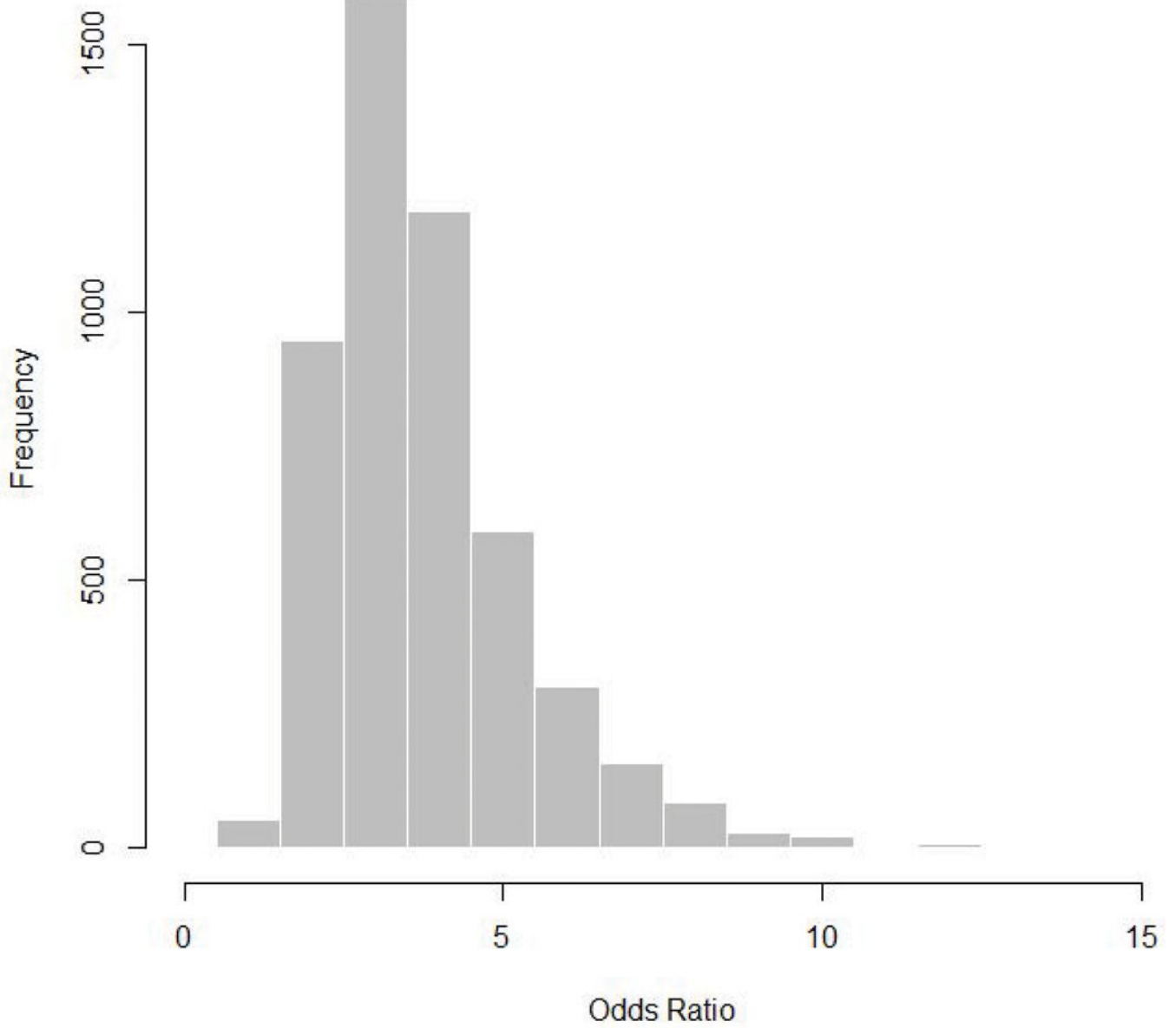


Figure 2. Posterior Distribution of Odds Ratio for Membership in the High-Risk Class for RelSZ versus NC

Table 1

Structure of the ZIP Mixture Model*

Component 1	Component k (k = 2, ..., m)
$Y_{ij} \sim \text{ZIP}(\lambda_{ij1}, \pi_{ij})$	$Y_{ij} \sim \text{Poisson}(\lambda_{ijk})$
$\log(\lambda_{ij1}) = \beta_1 + \beta \cdot x_{ij} + u_{ij}$	$\log(\lambda_{ijk}) = \beta_k + \beta \cdot x_{ij} + u_{ij}$
$\text{logit}(\pi_{ij}) = \psi \cdot x_{ij} + w_{ij}$	
$u_{ij} \sim N(0, \sigma_u^2), w_{ij} \sim N(0, \sigma_w^2)$	$u_{ij} \sim N(0, \sigma_u^2)$

* We require $\beta_1 < \beta_2 < \dots < \beta_m$ to ensure the identifiability of the model.

Table 2

Demographic Characteristics of the Sample

	ReISZ		NC	
	Mean	(SD)	Mean	(SD)
Age (yrs)	49.27	(15.94)	42.78	(15.37)
Estimated Verbal IQ*	107.86	(12.78)	106.33	(11.52)
Education (yrs)	15.21	(2.66)	15.31	(2.40)
SES**	2.25	(0.95)	2.16	(0.96)
GAS	74.98	(10.57)	76.59	(10.56)
% Male	38.15%		39.82%	
n	173		113	

* Estimated from the vocabulary subtest of the WAIS-R [52].

** Estimated from the Hollingshead scale [53], as revised by our group.

Table 3

Results of Model Fitting

Model	# of Risk Classes	Log Posterior Density	Log-Likelihood	# of Parameters	BIC
Poisson Regression	-	-818.4	-988.3	4	1999.22
ZIP Regression	-	-716.5	-832.2	6	1698.34
ZIP Mixture	1	-721.4	-854.4	5	1737.02
ZIP Mixture	2	-499.0	-642.0	8	1329.15
ZIP Mixture	3	-471.3	-640.6	11	1343.46

Table 4

Estimated Parameters (95% Posterior Intervals) for the Final Model

Parameter		Low-Risk Class (Component 1)	High-Risk Class (Component 2)
Poisson mean	Males	2.09 (1.66 to 2.57)	12.38 (10.54 to 14.43)
	Females	1.56 (1.15 to 2.04)	9.22 (7.58 to 11.26)
Rate of Zero-inflation		28.3% (17.3% to 37.8%)	—