Published in final edited form as:

Stat Med. 2015 July 30; 34(17): 2544–2558. doi:10.1002/sim.6496.

Positing, Fitting, and Selecting Regression Models for Pooled Biomarker Data

Emily M. Mitchell^{a,*}, Robert H. Lyles^b, and Enrique F. Schisterman^a

^aDivision of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland 20892, U.S.A

^bDepartment of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, U.S.A

Abstract

Pooling biospecimens prior to performing lab assays can help reduce lab costs, preserve specimens, and reduce information loss when subject to a limit of detection. Since many biomarkers measured in epidemiological studies are positive and right-skewed, proper analysis of pooled specimens requires special methods. In this paper, we develop and compare parametric regression models for skewed outcome data subject to pooling, including a novel parameterization of the gamma distribution that takes full advantage of the gamma summation property. We also develop a Monte Carlo approximation of Akaike's Information Criterion (AIC) applied to pooled data in order to guide model selection. Simulation studies and analysis of motivating data from the Collaborative Perinatal Project (CPP) suggest that using AIC to select the best parametric model can help ensure valid inference and promote estimate precision.

Keywords

AIC; Biomarkers; Gamma; MCEM; Pooled specimens; Skewness

1. Introduction

Estimating relationships between biomarkers and risk factors is an important aspect of epidemiology. Biomarkers can provide a useful surrogate when clinical endpoints are difficult to measure, or when symptoms take a long time to present [1]. In this study, we use specimens collected from the Collaborative Perinatal Project (CPP) to measure the association between potential risk factors and levels of the cytokine interferon gamma inducible protein (IP). This cytokine has recently been indicated as a potential marker for blood brain barrier dysfunction in patients with a rare autoinflammatory disease [2]. Often, however, biomarkers such as IP can be expensive to measure. By strategically pooling biospecimens prior to performing lab assays, the association between risk factors and

*Correspondence to: Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland 20892, U.S.A.

Supporting Information

Additional supporting information may be found in the online version of this article at the publishers web site.

biomarker levels can be estimated with a high level of precision, at only a fraction of the cost.

In the group testing scenario, pooling can effectively reduce cost while adequately determining which patients have a particular disease [3–5]. Similar to the original motivation for pooling popularized by Dorfman [6], the goal of these methods is often to reduce the total lab cost by minimizing the number of lab tests required to identify infected patients. Many epidemiological studies, on the other hand, strive to characterize population-level associations, rather than individual disease status. In such scenarios, pools can be assessed using special methods in regression formulations, where a pooled measurement represents the arithmetic mean of the biomarker concentrations of specimens comprising that pool [7–11]. Logistic regression models have been developed to estimate odds ratios when an outcome or an exposure variable is subject to pooling [12, 13]. When performing linear regression on a pooled outcome, pooling strategies have been developed to improve estimate efficiency [14–16], while reducing the total number of lab tests. In addition to reducing lab costs, pooling has the potential to mitigate information loss when biospecimens are subject to a limit of detection [8, 10, 17] and can help meet minimum volume requirements for certain lab tests by combining multiple specimens into a single sample [10, 17].

Regardless of the motivation for pooling, proper analysis of the resulting pooled measurements must be employed. Epidemiologists are often reluctant to pool specimens, in part because the necessary statistical methods are relatively new. In addition, few datasets exist that contain measurements on actual pooled samples, so that the efficacy of these analytical methods cannot be fully assessed via data analysis, but rather must rely heavily on simulation study results. Data from the CPP are unique in this regard since both individual-level as well as pooled specimens have been measured, permitting direct evaluation of our proposed methods.

The CPP followed eligible women from 1959 to 1974, gathering information on pregnancy outcomes, demographic information, and biological characteristics, including levels of circulating cytokines [18]. Many biomarkers, such as cytokine IP, can display a positive, right-skewed distribution, requiring the development of specialized analytical techniques [19–21]. When performing regression on skewed outcomes, a common approach is to assume a log-normal distribution in order to take advantage of the convenient properties of the normal distribution applied to a log transformation on the outcome. The gamma distribution, however, is also appropriate for modeling positive and continuous right-skewed outcomes, and is likewise often used in conjunction with a log link in generalized linear regression models. The gamma distribution can be particularly attractive for modeling data on pooled specimens, because if the individual-level measurements are gamma, the pooled measurements (assumed to be the mean of individual specimens) also follow a gamma distribution in certain cases. Recent epidemiological studies involving pooled specimens have successfully utilized this summation property in special cases of gamma regression [21, 22].

When pools are formed from specimens that do not share identical covariate values, the convenient summation property does not generally apply under the usual gamma regression parameterization. Nevertheless, when at most two covariate groups are represented in each pool, the observed likelihood for pooled outcome data can be maximized after expressing the density of each pool via a convolution integral [22]. For pools containing specimens with more than two unique covariate values, however, numerically evaluating the likelihood quickly becomes computationally intractable due to high-dimensional integrals required to specify the pooled density.

We consider several methods for dealing with this issue, which we then assess in simulation studies and apply to real data. First, we summarize a recently developed method which calculates MLEs based on the standard gamma or lognormal models by applying specific versions of the Monte Carlo Expectation Maximization (MCEM) algorithm [19]. Second, we propose use of an alternate parameterization for gamma regression, which can take full advantage of the gamma summation property for all types of pooled data. Next, we apply a Monte Carlo version of Akaike's Information Criterion (AIC) in order to improve precision of the resulting estimates by guiding selection of the best fitting parametric distribution. To our knowledge this is the first adaptation of Monte Carlo methods to estimate AIC in order to compare regression models for pooled outcome data. We use simulation studies to test the ability of AIC to choose the best model, and we apply each of these methods to estimate the relationship between risk factors and IP levels from the CPP dataset, comparing results from the individual-level data to those from pooled measurements. In addition to analyzing data from the CPP study, we also consider biomarker measurements from the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial. Although this study did not apply pooling techniques to the actual specimens, we analyze both individual and artificially pooled measurements of inhibin from this study in order to provide an additional illustration of the use of AIC to select the best-fitting model.

2. Regression Models

Suppose N subjects are separated into n groups (later to be defined as pools), where group i contains k_i subjects, so that $N = \sum_{i=1}^n k_i$, and let the ' ij ' subscript denote the j^{th} subject in the i^{th} group. Suppose that the individual-level outcome Y_{ij} is continuous and right-skewed, and is associated with its corresponding vector of covariates $\mathbf{x}_{ij}$ such that one of the following models holds:

$$\log(Y_{ij}) = \alpha + \mathbf{x}_{ij}\boldsymbol{\beta} + \varepsilon_{ij} \quad (1)$$

$$\log(\mu_{ij}) = \alpha + \mathbf{x}_{ij}\boldsymbol{\beta} \quad (2)$$

The first model (1) represents a linear regression on the log-transformed outcome under the assumption $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$. Although this common distributional assumption for the errors is not required for valid estimation of the coefficients, it is needed to compare model fit via likelihood-based criteria such as AIC. The second model (2) is a generalized linear

regression model with log link, such that $\mu_{ij} = E(Y_{ij})$ and Y_{ij} is assumed to follow a distribution in the exponential family.

2.1. MCEM Approaches to Maximum Likelihood

When individual-level data are known, standard software packages can often be applied to calculate estimates of the regression coefficients of interest (β) under these models. When only pooled data are available, however, direct optimization of the log-likelihood might be necessary. The log-likelihood for pooled outcomes is defined as:

$$l(\theta) = \sum_{i=1}^n \log f_p(Y_i^p | \mathbf{X}, \theta) \quad (3)$$

where $f_p(Y_i^p | \mathbf{X}, \theta)$ represents the density for pool i , and each Y_i^p represents the arithmetic mean of the individual specimens comprising that pool (i.e. $Y_i^p = k_i^{-1} \sum_{j=1}^{k_i} Y_{ij}$). In general, when pools are comprised of no more than two specimens, this density can be characterized by the single convolution integral,

$$f_p(Y_i^p | \mathbf{X}, \theta) = \int_0^{2Y_i^p} f(2Y_i^p - t | \mathbf{x}_{i1}, \theta) f(t | \mathbf{x}_{i2}, \theta) dt, \quad (4)$$

where $f(t | \mathbf{x}, \theta)$ represents the assumed distribution on the individual-level data that depends on the covariate vector $\mathbf{x}$ and parameter vector θ . In such cases, (3) can often be directly optimized by applying numerical integration techniques. Larger pool sizes, however, can introduce higher-dimensional integration, making direct optimization of the log-likelihood intractable. Recent research has successfully facilitated maximum likelihood (ML) estimation by developing a Monte Carlo Expectation Maximization (MCEM) algorithm specific to pooled data, applying importance sampling techniques to speed convergence [19]. These studies have shown that, for a log-normally distributed outcome and pools of size 2, the MLEs calculated under MCEM agree closely with those calculated by direct maximization of the log-likelihood via (4), generally yielding nearly identical estimates of coefficients and their standard errors.

The MCEM algorithm, similar to standard EM, seeks to maximize the expectation of the complete log-likelihood given the observed data [23]. At iteration $(t + 1)$ of the EM algorithm, parameter estimates are updated by maximizing $Q(\theta | \theta^t) = E_g\{l_c(\theta) | \mathbf{Y}^p, \mathbf{X}, \theta^t\}$ with respect to θ , where g indicates the density of the complete data conditional on the observed data, l_c denotes the complete log-likelihood, and $\mathbf{Y}^p = (Y_1^p, \dots, Y_n^p)$ represents the observed vector of pooled outcome measurements.

When a closed form of Q is difficult to achieve, we approximate this value using Monte Carlo methods. Let $E_g\{h(\mathbf{Y}_{i[-1]})\}$ represent any of the conditional expectations contained in Q for pool i , where $\mathbf{Y}_{i[-1]} = (Y_{i2}, \dots, Y_{ik_i})$ denotes the vector of missing data. Y_{i1} is omitted from the missing data since, given Y_i^p and $\mathbf{Y}_{i[-1]}$, its value is fixed. Applying importance sampling techniques to speed convergence [24], this value can be approximated by Monte Carlo methods:

$$E_g\{h(\mathbf{Y}_{i[-1]})\} \approx \frac{\sum_{m=1}^M h(\mathbf{Y}_{i[-1],m}) w(\mathbf{Y}_{i[-1],m})}{\sum_{m=1}^M w(\mathbf{Y}_{i[-1],m})}, \quad (5)$$

where $w(\mathbf{Y}_{i[-1]}) = g(\mathbf{Y}_{i[-1]} | Y_i^p, \mathbf{X}, \boldsymbol{\theta}^{(t)}) / g^*(\mathbf{Y}_{i[-1]} | Y_i^p, \mathbf{X}, \boldsymbol{\theta}^{(t)})$ represents the importance weights, and each $Y_{ij,m}$ is generated under $g^*(\mathbf{Y}_{i[-1]} | Y_i^p, \mathbf{X}, \boldsymbol{\theta}^{(t)})$, an alternate density that facilitates generation of samples.

Once MLEs have been calculated, standard error estimates may be approximated by applying similar MC techniques in conjunction with Louis's method [19, 25]. In subsequent simulation studies, we apply these MCEM methods to calculate MLEs when direct optimization of the observed log-likelihood is difficult.

While these MCEM techniques could theoretically be applied to any distribution, the form of the generating function and subsequent importance weights is specific to the assumed outcome density. In the following sections, we define these generating functions and corresponding weights for an outcome that follows a log-normal or gamma distribution.

2.2. Lognormal

While assuming a lognormal distribution on individual-level data can typically simplify analyses, it can induce additional complexity when only pooled measurements are available. In such cases, we recommend implementation of the MCEM method outlined in Section 2.1. For the MC approximation step, we first generate $Z_{ij,m}$ from the lognormal density $f(z_{ij,m} | \mathbf{x}_{ij}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma^{(t)})$, such that $E(\log z_{ij,m}) = \mu_{ij}^{(t)} = \boldsymbol{\alpha}^{(t)} + \mathbf{x}_{ij} \boldsymbol{\beta}^{(t)}$ and $\text{Var}(\log z_{ij,m}) = (\sigma^{(t)})^2$. We then define $Y_{ij,m} = k_i Y_i^p Z_{ij,m} / \sum_{j=1}^{k_i} Z_{ij,m}$, so that our new sample meets the linear constraint that the average of each pool is equivalent to the observed value for that pool (i.e.

$\sum_{j=1}^{k_i} Y_{ij,m} = k_i Y_i^p$). The importance weights to be used in (5) are then defined as:

$$w(\mathbf{Y}_{i[-1],m}) = \exp \left[- \left(2\sigma^{(t)2} k_i \right)^{-1} \left\{ \sum_{j=1}^{k_i} \left(\log y_{ij,m} - \mu_{ij}^{(t)} \right)^2 \right\} \right]$$

where $y_{i1,m} = k_i Y_i^p - \sum_{j=2}^{k_i} y_{ij,m}$. Additional details and examples concerning the implementation of this MCEM algorithm for lognormal regression with outcome pooling can be found in Mitchell *et al.* [19].

2.3. Gamma with Constant Shape

The gamma distribution is also common for modeling continuous right-skewed random variables, and its convenient summation properties can simplify analysis on pooled measurements [21]. Gamma regression is also closely associated with quasi-likelihood models, making it robust to distributional misspecification [26].

Suppose that $Y_{ij} \sim \text{Gamma}(a_{ij}, b_{ij})$, where a_{ij} and b_{ij} are the shape and scale parameters, respectively. Let $f(Y_{ij})$ denote the gamma density for the observation from the j^{th} subject in the i^{th} group, such that

$$f(Y_{ij}) = \{\Gamma(a_{ij})b_{ij}^{a_{ij}}\}^{-1} e^{-Y_{ij}/b_{ij}} Y_{ij}^{a_{ij}-1},$$

where $a_{ij}, b_{ij} > 0$, $\mu_{ij} = E(Y_{ij}) = a_{ij}b_{ij}$ and $\text{Var}(Y_{ij}) = a_{ij}b_{ij}^2$. In most standard GLM procedures (e.g., PROC GENMOD in SAS and *glm* in R), the default parameterization is to assume a constant shape parameter (a) and allow the scale parameter (b_{ij}) to model the expectation. Under a log link, $b_{ij} = a^{-1}\mu_{ij} = a^{-1} \exp(\alpha + \mathbf{x}_{ij}\boldsymbol{\beta})$. This parameterization maintains a constant coefficient of variation (CV), where $CV = \sqrt{\text{Var}(y)/E(Y)}$, a property shared by the lognormal distribution. We will refer to this parameterization as the ‘standard gamma’ model.

Although the standard gamma model permits the scale parameter to vary across all specimens, an $\mathbf{x}$ -homogeneous pooling strategy, where every specimen in a pool has identical covariate values, maintains a constant within-pool scale since all $\mathbf{x}_{ij} = \mathbf{x}_i$ and thus $b_{ij} = a^{-1} \exp(\alpha + \mathbf{x}_i\boldsymbol{\beta})$ for all $j = 1, \dots, k_i$. This property permits application of the gamma summation property, whereby $Y_i^p \sim \text{Gamma}(k_i a, b_i k_i^{-1})$ and the mean model based on pooled specimens is then:

$$\log(\mu_i) = \log\{E(Y_i^p)\} = \log \left\{ E \left(\frac{1}{k_i} \sum_{j=1}^{k_i} Y_{ij} \right) \right\} = \alpha + \mathbf{x}_i\boldsymbol{\beta},$$

Furthermore, since $\text{Var}(Y_i^p) = k_i^{-1}(\mu_i^2/a)$, a weighted regression with weights $\{k_i : i = 1, \dots, n\}$ can easily be applied using standard software. Alternatively, the observed data likelihood can be maximized directly using optimization routines such as PROC NLMIXED in SAS or *optim* in R.

An advantage of fitting the gamma with constant shape model to $\mathbf{x}$ -homogeneous pools is that it produces identical coefficient estimates as under the full data (see Appendix A.1 for proof). While it may be tempting to pool all specimens with identical covariate values, doing so will almost certainly result in poor standard error estimates. Thus, it is recommended to form enough pools so that resulting inference based on both estimates and standard errors can be trusted [9].

When pools are $\mathbf{x}$ -heterogeneous, the summation property of the gamma distribution no longer applies under the standard gamma model. Instead, we can use the MCEM methods outlined in Section 2.1. This method is similar to that applied to the lognormal distribution. Aside from replacing any lognormal densities with gamma densities, the main difference is the calculation of the importance weights. When the standard gamma model is assumed, the $Z_{ij,m}$ are generated under the marginal gamma density, and the $Y_{ij,m}$ are defined to meet the

linear constraint. The importance weights corresponding to the gamma density to be utilized in (5) are:

$$w(\mathbf{Y}_{i[-1],m}) = c(Y_i^p) \left(\sum_{j=1}^{k_i} \frac{ay_{ij,m}}{\mu_{ij}^{(t)}} \right)^{ak_i} \exp \left(- \sum_{j=1}^{k_i} \frac{ay_{ij,m}}{\mu_{ij}^{(t)}} \right),$$

where $y_{i1,m} = k_i Y_i^p - \sum_{j=2}^{k_i} y_{ij,m}$, $\mu_{ij}^{(t)} = \exp(\alpha^{(t)} + \mathbf{x}_{ij}\boldsymbol{\beta}^{(t)})$, and $c(\bullet)$ is a function of Y_i^p not involving the parameters of interest.

2.4. Gamma with Constant Scale

In addition to the standard gamma model, we consider a second parameterization, which we will refer to as the ‘alternate gamma’ model. This parameterization assumes a constant scale (b) and models the mean as a function of the shape parameter (a_{ij}). Here, we set $a_{ij} = b^{-1}\mu_{ij} = b^{-1} \exp(\alpha + \mathbf{x}_{ij}\boldsymbol{\beta})$, and the relationship between the expectation and variance is now linear, such that $\text{Var}(Y_{ij}) = b\mu_{ij}$. Although this parameterization of the gamma distribution may not be as common, it has the advantage of simplifying analysis of pooled specimens. Note that this parameterization of the gamma distribution as well as the previous one both model the mean as $\mu_{ij} = \exp(\alpha + \mathbf{x}_{ij}\boldsymbol{\beta})$. The difference between these two parameterizations is determined by which parameter (shape or scale) is held constant.

For individual level data, MLEs are calculated by maximizing the log-likelihood with respect to $(\boldsymbol{\beta}^*, b) = (\alpha, \boldsymbol{\beta}, b)$, where

$$l(\boldsymbol{\beta}^*, b) = \sum_{i=1}^n \sum_{j=1}^{k_i} \{ -\log \Gamma(a_{ij}) - a_{ij} \log b - Y_{ij}/b + (a_{ij} - 1) \log Y_{ij} \}, \quad (6)$$

(sample code is provided in Appendix B). This model can prove particularly appealing when the outcome measurement is on pooled specimens, as the constant scale parameter allows exploitation of the gamma sum property regardless of the homogeneity of pools. It follows that the pooled values maintain a gamma distribution, such that

$Y_i^p \sim \text{Gamma}(\sum_{j=1}^{k_i} a_{ij}, bk_i^{-1})$. Maximizing the log-likelihood for pooled data is straightforward, requiring only a slight variation from (6), where we now maximize:

$$l(\boldsymbol{\beta}^*, b) = \sum_{i=1}^n \{ -\log \Gamma(a_i) - a_i \log(bk_i^{-1}) - k_i Y_i^p / b + (a_i - 1) \log T_i^p \} \quad (7)$$

where $a_i = \sum_{j=1}^{k_i} a_{ij} = b^{-1} \sum_{j=1}^{k_i} \exp(\alpha + \mathbf{x}_{ij}\boldsymbol{\beta})$. Under $\mathbf{x}$ -homogeneous pools, a_i reduces to $a_i = k_i b^{-1} \exp(\alpha + \mathbf{x}_i\boldsymbol{\beta})$. Again, numerical optimization methods can be readily used to maximize (7) with respect to b and $\boldsymbol{\beta}^*$. This alternate parameterization provides an accessible and defensible procedure to analyze pooled, right-skewed outcomes while avoiding the computational complexity of the MCEM algorithm, and it is particularly suited to data that exhibit a constant mean-variance relationship.

3. Model Selection

Now that we have defined several models and methods to calculate MLEs when an outcome is pooled, we focus on tools to help choose the best model. Model misspecification is often a concern when fitting parametric models, and choosing the distribution that fits the data best can help optimize precision. To guide selection of the best model, we develop a Monte Carlo approximation to Akaike's Information Criterion (AIC) to evaluate goodness of fit of the models defined in the previous sections. While the log-normal and standard gamma regression models are often interchangeable [27], AIC can provide a convenient and useful guide for selecting the best parametric model and has been shown to effectively distinguish between the lognormal and gamma models for individual-level (non-pooled) data [28, 29].

While each of the aforementioned models may be well suited to model a positive, right-skewed outcome, they are unlikely to fit equally well to a particular set of data. A constant mean-variance relationship, for instance, should be best fit by the alternate gamma model, whereas an outcome exhibiting a constant CV would be better modeled by the lognormal or standard gamma models. Figure 1 illustrates how individual-level data generated under each of these models may differ. Each of the models are generated with linear predictor $\eta_i = -1 + x_i$, where $x_i \sim \text{Bernoulli}(0.5)$ for $i = 1, \dots, 10000$. The first row shows histograms of data generated under a lognormal distribution, such that $\log(Y_i) \sim N(\eta_i, 0.5^2)$. The second row is from data generated under a standard gamma model, with $Y_i \sim \text{Gamma}(a = 2, b_i = e^{\eta_i}/2)$, and the final row has $Y_i \sim \text{Gamma}(a_i = 2e^{\eta_i}, b = 1/2)$. The right column provides the histograms of the log of the outcome, separated by the levels of x . As illustrated by Figure 1, a log-transformation on the outcome (i.e. lognormal model) may provide a reasonable fit for data generated under a standard gamma model, but not under the alternate gamma model. Similarly, the gamma model with constant scale (alternate parameterization) may not be appropriate to model data from the lognormal or standard gamma distributions. As demonstrated later in the Data Analysis (Section 4), while each of these distributions may provide similar parameter estimates, choosing the best-fitting distribution can improve estimate precision and, consequently, power to detect significant associations.

In order to choose the best model among these three alternatives, we apply an AIC-based assessment to both full and pooled data to compare models under different distributional assumptions. Details concerning the use of AIC to select between models with differing probability distributions can be found in [28]. For full data or a random sample of individual specimens, AIC is defined as:

$$AIC = -2l(\boldsymbol{\theta}) + 2K = -2 \sum_{i=1}^n \sum_{j=1}^{k_i} \log f(Y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\theta}) + 2K,$$

where K refers to the total number of estimated parameters in the model, and a lower value of AIC indicates better fit [30]. An important note is that in order to compare these values between the gamma and lognormal models, the data must be on the same scale. For instance, AIC calculated under the lognormal distribution should be composed of the lognormal density, not the normal density on the log-transformed outcome. In this scenario, we are

interested in determining the best-fitting parametric distribution, assuming a fixed collection of covariates. Thus, application of the AIC will essentially be equivalent to a comparison of the observed data likelihood under the lognormal and gamma distributions, for instance. Similarly, additional information criteria such as AICc or BIC will essentially be equivalent, since their distinguishing characteristics (e.g. number of parameters, sample size) are identical for all tested models. In the remainder of this paper, we will continue to refer to this likelihood-based selection method as ‘AIC’ for simplicity and convenience.

For standard likelihood-based regression methods, calculation of AIC is straightforward. The proposed MCEM algorithms for pooled outcome data, however, do not directly provide an observed data likelihood. Instead, we propose and demonstrate the approximation of AIC using additional Monte Carlo methods.

When available, the value of the maximized observed log-likelihood can be inserted directly into the AIC equation. For models that employ the MCEM algorithm to calculate MLEs, such as the lognormal model applied to pooled outcomes or the standard gamma model fit to heterogeneous pools, we recommend a Monte Carlo estimation of the observed log-likelihood. For these models, $f_p(Y_i^p)$ can be approximated by:

$$f_p(Y_i^p) \approx \frac{k_i}{M} \sum_{m=1}^M \left\{ f \left(k_i Y_i^p - \sum_{j=2}^{k_i} Y_{ij,m} | \mathbf{x}_{i1}, \hat{\boldsymbol{\theta}} \right) I \left(\sum_{j=2}^{k_i} Y_{ij,m} < k_i Y_i^p \right) \right\} \quad (8)$$

where each $Y_{ij,m}$ is generated from $f(y_{ij} | \mathbf{x}_{ij}, \hat{\boldsymbol{\theta}})$ for $m = 1, \dots, M$, and $\hat{\boldsymbol{\theta}}$ denotes the MLE (derivation provided in Appendix A.2). A similar concept concerning this type of MC estimation is applied in Dupuis *et al.* [31]. By the law of large numbers, this approximation will converge to $f_p(Y_i^p)$ for large M . For the simulation studies and data analysis in this paper, we set $M = 10,000$. An alternative information criterion specific to the EM algorithm is given by Ibrahim *et al.* [32], in which the information criterion is based on the Q function of the EM algorithm and is an immediate by-product of the MCEM process [33]. While this information criterion provides a convenient method to compare two regression models with MLE’s calculated via MCEM, more complex strategies are needed to approximate the AIC based on the observed likelihood, requiring both the ‘Q’ and ‘H’ functions in traditional EM notation [32]. In general, we found the Monte Carlo approximation given in (8) to be both a straightforward and effective way to obtain an AIC value that can be directly compared with AICs derived from direct optimization strategies, such as that arising from the Gamma distribution with constant scale parameterization.

4. Simulation Study

In this section, we perform simulation studies to assess the effects of using AIC as the only criterion for model selection. In particular, we compare the performance of estimates from AIC-selected models to those fit under the true distribution. For each of the simulation studies, 5000 simulations were performed in R. Datasets were generated to resemble the CPP dataset, treating interferon gamma inducible protein (IP) as the outcome. Predictor variables were simulated to mimic the covariates where X_1 = age (years), X_2 = smoking status (yes/no), X_3 = race (1 = white/2 = black), and X_4 = spontaneous abortion (SA) status

(yes/no). X_1 was generated from a normal distribution with a mean of 27 and standard deviation of 6.5, then rounded to the nearest whole number. X_2 , X_3 , and X_4 were each generated from Bernoulli distributions with probabilities 0.47, 0.30, and 0.45. The simulated outcome (Y) was based on a generalized linear model fit to the individual-level IP data, such that

$$\eta = -4.3 + 0.01(\text{Age}) + 0.02(\text{Smoking Status}) + 0.17(\text{Race}) - 0.1(\text{SA status}).$$

For all simulations, three different sets of outcomes with sample size $N = 650$ were generated under the following distributions:

1. $\log(Y) \sim N(\eta, 1)$
2. $Y \sim \text{gamma}(1.24, \exp(\eta)/1.24)$ (standard parameterization)
3. $Y \sim \text{gamma}(\exp(\eta)/0.02, 0.02)$ (alternate parameterization)

In each simulation, models were fit under the full data, $\mathbf{x}$ -homogeneous pools, and heterogeneous pools. 400 $\mathbf{x}$ -homogeneous pools were formed to have similar pool sizes, in an effort to mimic the number of pools in the data analysis. To test the effect of larger pool sizes, 150 heterogeneous pools were formed based on a k -means clustering strategy to promote efficiency [19]. For the full data, the *glm* function in R was applied to the lognormal and standard gamma models.

For the first simulation, the outcome is generated from the lognormal distribution specified above. The next simulation treats data generated from the standard gamma distribution, and the final simulation generates outcomes from the alternate gamma model. At each simulation, we calculate AIC values under each of the proposed models, applying the Monte Carlo approximation when necessary. Results from these models chosen by AIC are provided in order to assess performance when the parametric distribution is selected based on this criterion.

Table 1 shows simulation results for the frequency of model selection for data generated and fit under each of the three distributions: (1) lognormal (2) standard gamma and (3) alternate gamma, for various pooling types, where $\mathbf{x}$ -homogeneous and heterogeneous pools were formed to promote estimate precision by combining specimens with similar covariate values. These results are also compared to the performance of AIC applied to a random selection of the same size (i.e. same number of lab assays). Simulation conditions are described in detail in Section 4. As indicated by this simulation study, AIC arguably provides an effective measure for selecting the best distribution, correctly choosing the lognormal distribution over 95% of the time, and accurately identifying the best gamma model over 60% of the time. Furthermore, our MC approximation techniques tend to perform as well as the closed-form AIC under a random sample of size equal to the number of pools, suggesting that the ability of AIC to select the best model was not hindered by pooling nor by the additional MC methods required to estimate AIC when MCEM was necessary to calculate MLEs. The discriminatory ability of AIC, in general, will be largely data specific. In this simulation study, the CPP data example that motivated the design of the

simulations had similar AIC values for both of the Gamma models, a characteristic that was reflected in the discriminatory capacity of AIC in the simulation results. However, a different dataset with an alternate distribution of the biomarker of interest may demonstrate more similarity between the lognormal and standard gamma models, for instance, with a higher rate of separation between the two gamma models.

4.1. Model Validity

To assess the validity of these models, mean bias, ratio of estimated standard error ($\hat{SE}$) to empirical standard deviation (SD), and 95% CI coverage are provided in Table 2, where a valid estimation procedure will exhibit a mean bias close to 0, nearly nominal 95%

confidence interval (CI) coverage, and $\hat{SE}/SD$ close to 1. The simulations presented here compare the model under the correctly-specified distribution with that from the model selected based on the lowest AIC value. Additional simulation results for all models are provided in the Supplementary Materials in order to illustrate the potential repercussions of distributional misspecification.

These simulations demonstrate that when AIC is applied to select the best distribution, model estimates remain essentially unbiased, closely approximating results from the true model. Standard error estimates, while slightly underestimated for the AIC-selected model, still maintain a high level of accuracy, with confidence interval coverage dropping only slightly below 95%.

Perhaps even more compelling is the similarity between estimates calculated under data containing pooled measurements and those from the complete set of individual-level measurements ('Full Data'). These results demonstrate that the proposed analytical methods for parameter estimation and model selection for pooled data can effectively extract the information contained in the full dataset with respect to unbiased estimation and confidence interval coverage.

4.2. Precision

Now that we have demonstrated the ability of the MC AIC to select regression models which provide accurate parameter estimates, we compare efficiency under the true model with that under the model selected via AIC. Table 3 provides results on the estimate precision of each model under the three different types of pooling, where lower empirical standard deviation (SD) indicates a more precise estimation procedure. With respect to efficiency, the model under the true distribution, as expected, produces the most precise estimates (lowest SD), for all types of pooling. Using AIC to select the best model, however, does extremely well, nearly reaching the optimal precision levels of the true model. Note that precision values under the standard gamma model are identical for full data and $\mathbf{x}$ -homogeneous pools, as a result of the MLE preservation property mentioned in Section 2.3. This high level of precision is maintained for the standard gamma model under heterogeneous pools, given the informative pooling strategy applied via k -means clustering.

In this simulation study, AIC performs well when discriminating between the lognormal and gamma distributions (as reflected in Table 1), resulting in ideal estimates under the AIC-

based model, particularly when the outcome is generated from a lognormal distribution. If this were not the case, and instead it was difficult to distinguish between the lognormal and gamma models, then the two distributions would likely fit the data equally well, and the results from the AIC-based model would still provide valid and efficient estimates of the regression coefficients.

5. Data Analysis

5.1. Cytokines in the CPP

After removing 6 outliers, 645 individual-level measurements from the CPP dataset were available for analysis. In addition to measuring IP values on individual specimens, 496 of those specimens had been physically combined and measured in pools of 2, formed homogeneously with respect to participant spontaneous abortion (SA) status [21]. This subsequent pooling produced a secondary, hybrid dataset consisting of the 248 pooled values as well as the initial measurements on 149 of the individual specimens that were not pooled. In addition to these observed pooled values, we also artificially recreate these pools based on the individual measurements, in order to obtain the regression estimates on the expected values of the pools, calculated as the mean of the IP values from the specimens comprising each pool. Although the pooled concentration generally is taken as an average of the individual biomarker concentrations comprising that pool, assay measurement error can affect both pooled and individual measurements, creating apparent discrepancy in the observed and expected pooled values [9]. While correcting for this source of error is beyond the scope of this paper, results for both the expected and observed pools are provided to emphasize the potential impact of assay measurement error on biomarker analysis. Each of the resulting datasets, (full data, observed pools, and expected pools) were fit to each of the available models. AIC values, where applicable, were generated based on the log-likelihood of the pooled data. For models fit under the MCEM method, the Monte Carlo approximation of the pooled log-likelihood was estimated with an MC size of 10,000.

The variables age, smoking status, race, and spontaneous abortion were chosen as covariates in the regression model as these factors have previously been identified as having a possible association with cytokine levels [18, 34, 35]. As illustrated in Table 4, regression on the full, non-pooled dataset identifies a statistically significant association between race and IP levels under all distributional assumptions. While different distributions result in different parameter estimates, AIC values on the full data as well as the pooled data clearly select the lognormal as the most appropriate distributional assumption for this dataset. Even when specimens are pooled, estimates are similar to those calculated for the full data, and inference remains the same. Namely, that the lognormal distribution is the best fit, and that, based on this distributional assumption, race is significantly associated with levels of cytokine IP at the 0.05 level. This consistency is reassuring, suggesting that the same results would have been found even if only measurements on the pooled specimens had been available.

5.2. Inhibin in the EAGeR Trial

As an additional illustration of these methods, we consider levels of the protein inhibin, measured as part of the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial. This study enrolled 1228 women in an effort to determine the effect of aspirin on pregnancy outcomes. As part of this study, blood specimens were collected and measured for various biomarkers, including the protein inhibin. For this example, we treat inhibin as the outcome in a regression setting, and perform regression on both the individual specimens as well as artificially pooled specimens. Since inhibin levels may be predicted by pregnancy-related variables, covariates included in the analysis include number of previous pregnancy losses and anti-mullerian hormone (AMH) levels. In addition, education status ($>$ High School) was included as a potential indicator of stress, and an indicator of a close family member with auto-immune disease was included since proclivity towards autoimmune disease has been implicated in fertility-related outcomes [36]. After removing any observations with missing data, 1142 observations remained. In order to assess the results under pooled specimens, 500 pools were artificially formed based on a k -means clustering strategy in order to improve precision. Resulting pools contained 1 to 8 specimens.

Table 5 provides point estimates and standard errors for regression on the full dataset of individual specimens ($N = 1142$) as well as 500 artificially-formed pools. Since actual pools were not formed from the specimens in this study, comparison of the observed and expected pools is not possible. This example, however, illustrates a situation when the alternate gamma model provides the best fit to the data, indicated by the lowest AIC value under this distribution for both the individual-level specimens as well as an analysis on the pooled specimens. Parameter estimates are similar between each of the three models, although the alternate gamma enjoys slightly more precision, as evidenced by the lower standard error estimates.

6. Discussion

In this study, we developed methods to analyze three parametric models for data that are skewed under lognormal, standard gamma, and alternate gamma distributions. We found that using our proposed Monte Carlo AIC to distinguish between these distributions proved an effective tool to obtain valid inference by choosing the model best suited to the data. Even when AIC does not select the true distribution, overall results remain approximately unbiased and precision is only slightly reduced. Although we focused on applying AIC to choose between distributions, our proposed AIC methods could also be used to help select the best set of covariates in a model based on pooled outcome data. As pointed out by a reviewer, care must be taken when using AIC to choose the best set of covariate values, since a large number of candidate covariates could cause the AIC to fail. While this issue did not present itself in the current study, it is an important consideration that should be kept in mind when applying AIC to such models. Alternatives to AIC, such as BIC, AICc, or RIC, may also prove beneficial in helping choose between sets of covariates [28].

An alternative to selecting fully parametric models may be to apply quasi-likelihood regression, which requires fewer assumptions on the outcome. Ongoing research indicates that application of a full quasi-likelihood model would provide an additional layer of

robustness, particularly with respect to estimation of standard error estimates. This flexibility of quasi-likelihood models, however, is commonly accompanied by a loss of precision compared with correct specification of a parametric regression model. In addition, likelihood-based model selection criteria such as AIC are unavailable to compare quasi-likelihood models. Another alternative involves application of the Central Limit Theorem when pool sizes are large enough. Prior work on this topic has previously been considered by Caudill [37], where pools containing fewer than 7 specimens were shown to poorly approximate a normal distribution. If pool size is small or varied, as in the CPP analysis, this technique may not apply.

As illustrated in the CPP data analysis, real datasets are rarely immune to complications such as pooling and measurement error. In particular, this dataset exhibited differences in parameter estimates between the observed and expected pools, suggesting that further information concerning the impact of pooling and measurement error may influence not only the decision on whether to pool data, but also the treatment of that pooled data. While recent studies have researched these topics with respect to pooling [8, 9], additional investigation extending these ideas specifically to right-skewed, pooled outcomes, may prove helpful in order to maximize the accessibility and benefit of pooling biospecimens.

While pooling is almost always accompanied by at least some degree of information loss, the analytical methods proposed here can effectively mimic the results that would have been available if the entire set of individual-level specimens had been measured. If pools are strategically formed via homogeneous pools or a k -means pooling strategy, estimate precision from analysis on the pooled specimens can closely approximate the precision from analyzing all individual specimens. Thus, when pools are formed strategically and analyzed appropriately, valid inference can be preserved when a right-skewed biomarker is measured on pooled specimens.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Special thanks to Neil Perkins, Amita Manatunga, Qi Long, Michelle Danaher, and Brian Whitcomb for their insightful advice. This work was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health; and by the Long-Range Research Initiative of the American Chemistry Council. Additional support from the National Institute of Nursing Research [1RC4NR012527-01]; the National Institute of Environmental Health Sciences [5R01ES012458-07]; and the National Center for Advancing Translational Sciences of the National Institutes of Health [Award Number UL1TR000454] is also acknowledged. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Aronson JK. Biomarkers and surrogate endpoints. *British Journal of Clinical Pharmacology*. 2005; 59:491–494. [PubMed: 15842546]
2. Rodriguez-Smith J, Lin Y, Kim H, Montealegre Sanchez GA, Chapelle DC, Plass N, Tsai W, Huang Y, Gadina M, Bielekova B, Goldbach-Mansky R. A173: Cerebrospinal Fluid Cytokines Correlate

With Innate Immune Cells in Neonatal Onset Multisystem Inflammatory Disease (NOMID) Patients in Clinical Remission Treated With Anakinra. *Arthritis and Rheumatology*. 2014; 66:S226.

3. Delaigle A, Hall P. Nonparametric Regression with Homogeneous Group Testing Data. *The Annals of Statistics*. 2012; 40:131–158.
4. Tebbs JM, McMahan CS, Bilder CR. Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics*. 2011; 69:1065–1073.
5. Zhang B, Bilder CR, Tebbs JM. Regression analysis for multiple-disease group testing data. *Statistics in Medicine*. 2013; 32:4954–4966. [PubMed: 23703944]
6. Dorfman R. The detection of defective members of large populations. *The Annals of Mathematical Statistics*. 1943; 14:436–440.
7. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine*. 2003; 22:2515–2527. [PubMed: 12872306]
8. Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Pediatric and Perinatal Epidemiology*. 2008; 22:486–496. [PubMed: 18782255]
9. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 29:597–613. [PubMed: 20049693]
10. Schisterman EF, Vexler A, Ye A, Perkins NJ. A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity. *Annals of Applied Statistics*. 2011; 5:2651–2667.
11. Heffernan AL, Aylward LL, Toms LL, Sly PD, Macleod M, Mueller JF. Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology*. 2014; 24:225–232. [PubMed: 24192659]
12. Vansteelandt S, Goetghebuer E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. [PubMed: 11129470]
13. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
14. Ma CX, Vexler A, Schisterman EF, Tian L. Cost-efficient designs based on linearly associated biomarkers. *Journal of Applied Statistics*. 2011; 38:2739–2750.
15. Malinovsky Y, Albert PS, Schisterman EF. Pooling Designs for Outcomes under a Gaussian Random Effects Model. *Biometrics*. 2012; 68:45–52. [PubMed: 21981372]
16. Mitchell EM, Lyles RH, Manatunga AK, Perkins NJ, Schisterman EF. A highly efficient design strategy for regression with outcome pooling. *Statistics in Medicine*. 2014; 33:6305.
17. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics*. 2006; 7:585–598. [PubMed: 16531470]
18. Whitcomb BW, Schisterman EF, Klebanoff MA, Baumgarten M, Luo X, Chellini N. Circulating levels of cytokines during pregnancy: thrombopoietin is elevated in miscarriage. *Fertility and Sterility*. 2008; 89:1795–1802. [PubMed: 17706203]
19. Mitchell EM, Lyles RH, Manatunga AK, Danaher M, Perkins NJ, Schisterman EF. Regression for skewed biomarker outcomes subject to pooling. *Biometrics*. 2014; 70:202–211. [PubMed: 24521420]
20. Caudill SP. Important issues related to using pooled samples for environmental chemical biomonitoring. *Statistics in Medicine*. 2011; 30:515–521. [PubMed: 21312217]
21. Whitcomb BW, Perkins NJ, Zhang Z, Ye A, Lyles RH. Assessment of skewed exposure in case-control studies with pooling. *Statistics in Medicine*. 2012; 31:2461–2472. [PubMed: 22437722]
22. Perkins, NJ.; Whitcomb, BW.; Lyles, R.; Schisterman, EF. Flexible Assessment of Skewed Exposure in Case-Control Studies with Case-Specific and Random Pooling. Poster session (#49) presented at. Joint Statistical Meetings; 2011; July 30–August 4; Miami, FL.
23. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 1977; 39:1–38.

24. Lange, K. Numerical Analysis for Statisticians. 2. Springer; New York: 2010. Independent Monte Carlo.
25. Louis TA. Finding the Observed Information Matrix when Using the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1982; 44:226–233.
26. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 1974; 61:439–447.
27. Firth D. Multiplicative Errors: Log-Normal or Gamma? Journal of the Royal Statistical Society Series B (Methodological). 1988; 50:266–268.
28. Burnham, KP.; Anderson, DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2. New York: Springer; 2002. Selection When Probability Distributions Differ by Model.
29. Dick EJ. Beyond 'lognormal versus gamma': discrimination among error distributions for generalized linear models. Fisheries Research. 2004; 70:351–366.
30. Akaike H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.
31. Dupuis P, Leder K, Wang H. Importance sampling for sums of random variables with regularly varying tails. ACM Transactions on Modeling and Computer Simulation. 2007; 17:1–21.
32. Ibrahim JG, Zhu H, Tang N. Model Selection Criteria for Missing-Data Problems Using the EM Algorithm. Journal of the American Statistical Association. 2008; 103:1648–1658. [PubMed: 19693282]
33. Zhang B, Chen Z, Albert PS. Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. Biostatistics. 2012; 13:74–88. [PubMed: 21908867]
34. Cox ED, Hoffmann SC, DiMercurio BS, Wesley RA, Harlan DM, Kirk AD, Blair PJ. Cytokine Polymorphic Analyses Indicate Ethnic Differences in the Allelic Distribution of Interleukin-2 and Interleukin-61. Transplantation. 2001; 72:720–726. [PubMed: 11544437]
35. Lieberman JA, Moscicki AB, Sumerel JL, Ma Y, Scott ME. Determination of cytokine protein levels in cervical mucus samples from young women by a multiplex immunoassay method and assessment of correlates. Clinical and Vaccine Immunology. 2008; 15:49–54. [PubMed: 17978011]
36. Poppe K, Velkeniers B, Glinde D. The role of thyroid autoimmunity in fertility and pregnancy. Nature Clinical Practice Endocrinology and Metabolism. 2008; 4:394–405.
37. Caudill SP. Characterizing populations of individuals using pooled samples. Journal of Exposure Science and Environmental Epidemiology. 2010; 20:29–37. [PubMed: 19002216]

A. Proofs

A.1. Proof of preserved MLEs under x-homogeneous pools for standard gamma model

Suppose that all pools are formed homogeneously with respect to covariate values such that $Y_{ij} \sim \text{Gamma}(a, b_i)$ and $b_i = a^{-1}\mu_i = a^{-1} \exp(a + \mathbf{x}_i\boldsymbol{\beta})$ for all $j = 1, \dots, k_i$. Then the MLEs for $\boldsymbol{\beta}^* = (a, \boldsymbol{\beta})'$ under the full, individual-level data maximize the unpooled log-likelihood:

$$\begin{aligned}
 l_f(\boldsymbol{\beta}^*) &= \sum_{i=1}^n \sum_{j=1}^{k_i} \{ -\log \Gamma(a) - a \log(b_i) - Y_{ij}/b_i + (a-1) \log Y_{ij} \} \\
 &= c(a, \mathbf{Y}) - a \sum_{i=1}^n \sum_{j=1}^{k_i} \{ \alpha + \mathbf{x}_i \boldsymbol{\beta} + Y_{ij} e^{-(\alpha + \mathbf{x}_i \boldsymbol{\beta})} \} \\
 &= c(a, \mathbf{Y}) - a \sum_{i=1}^n \left\{ k_i (\alpha + \mathbf{x}_i \boldsymbol{\beta}) + e^{-(\alpha + \mathbf{x}_i \boldsymbol{\beta})} k_i Y_i^p \right\}
 \end{aligned}$$

where $c(a, \mathbf{Y})$ does not depend on β^* . For $\mathbf{x}$ -homogeneous pools, $Y_i^p \sim \text{Gamma}(k_i a, b_i k_i^{-1})$ for each pool, so that the MLEs for β^* based on the observed, pooled data maximize:

$$\begin{aligned} l_p(\beta^*) &= c(a, \mathbf{k}, \mathbf{Y}) + \sum_{i=1}^n \{ -\log \Gamma(k_i a) - k_i a \log(b_i / k_i) - k_i Y_i^p / b_i + (k_i a - 1) \log Y_i^p \} \\ &= c(a, \mathbf{k}, \mathbf{Y}) - a \sum_{i=1}^n \left\{ k_i (\alpha + \mathbf{x}_i \beta) + e^{-(\alpha + \mathbf{x}_i \beta)} k_i Y_i^p \right\} \end{aligned}$$

where, again, $c(a, \mathbf{k}, \mathbf{Y})$ does not depend on β^* . Thus, since the components of the log-likelihoods that depend on β^* are identical, the MLEs for β^* calculated from both the full and homogeneously pooled log-likelihood will also be identical.

A.2. Proof of MC AIC

For models that employ the MCEM algorithm to calculate MLEs, such as the lognormal model applied to pooled outcomes or the standard gamma model fit to heterogeneous pools, we recommend a Monte Carlo estimation of the observed log-likelihood into order to calculate AIC values. For these models, $f(Y_i^p)$ can be approximated by:

$$\begin{aligned} f(Y_i^p) &= \int \cdots \int_{Y_{ik_i}} f(Y_i^p, Y_{i2}, \dots, Y_{ik_i}) d\mathbf{Y}_{i[-1]} \\ &= \int \cdots \int_{Y_{ik_i}} k_i f \left(k_i Y_i^p - \sum_{j=2}^{k_i} T_{ij} | \mathbf{x}_{i1}, \boldsymbol{\theta} \right) I \left(\sum_{j=2}^{k_i} Y_{ij} < k_i Y_i^p \right) \prod_{j=2}^{k_i} f(Y_{ij} | \mathbf{x}_{ij}, \boldsymbol{\theta}) d\mathbf{Y}_{i[-1]} \quad (9) \\ &= E_{\mathbf{Y}_{i[-1]}} \left\{ k_i f \left(k_i Y_i^p - \sum_{j=2}^{k_i} T_{ij} | \mathbf{x}_{i1}, \boldsymbol{\theta} \right) I \left(\sum_{j=2}^{k_i} Y_{ij} < k_i Y_i^p \right) \right\}, \end{aligned}$$

Estimating this value with Monte Carlo methods is straightforward once the MLEs have been calculated, as we can approximate (9) with Equation (8) where each $Y_{ij,m}$ is generated from $f(y_{ij} | \mathbf{x}_{ij}, \hat{\boldsymbol{\theta}})$ for $m = 1, \dots, M$, and $\hat{\boldsymbol{\theta}}$ denotes the MLE.

B. Example R Code

R code to calculate MLEs from alternate gamma model. Code as shown applies to individual-level data as well as both $\mathbf{x}$ -homogeneous and heterogeneous pools. The defined *AVE* function alters the built-in R *ave* function so that it can be used in conjunction with *apply* without duplicating the function argument. The *Poolit* function was written to simplify the pooling process, by averaging (or summing) the data by cluster, then ordering by cluster number and dropping duplicates (i.e. keeping only one observation per pool). When optimizing the *gamma2* function, the 'y' vector should be a vector of unique pooled measurements, in the order of cluster number. In the *optim* statement, 'par' is a vector of starting values for $\theta = (\beta, \phi)$. In this code, the intercept (α) is included in the coefficient vector (β).

```
AVE = function(x, ..., fun = mean) ave(x, ..., FUN=fun)
Poolit = function(data, cluster, sums=F){
```

```
if(sums) fun = sum else fun = mean
unique(apply(cbind(cluster, data),2,
FUN=AVE, cluster, fun=fun)[order(cluster),,1][-1]
)
gamma2 = function(theta, y, X, cluster=1:length(y)){
b = theta[length(theta)]
Beta = theta[-length(theta)]
eta = as.matrix(cbind(1, X))%*%Beta
mu.ij = exp(eta)
mu.i = Poolit(mu.ij, cluster, sum=T)
a.i = mu.i/b
ki = tabulate(cluster)[tabulate(cluster)>0]
b.i = b/ki
ll = sum(dgamma(y, shape=a.i, scale=b.i, log=T))
return(-ll)
}
optim(par, fn=gamma2, y=y, X=X, cluster=cluster,
hessian=T, method="L-BFGS-B", lower=c(rep(-Inf, length(par)-1),1E-7))
```

Figure 1. Histograms of data generated under lognormal, standard gamma, and alternate gamma distributions.

Frequency of model selection (%) based on AIC. ‘Standard’ refers to the standard parameterization of the gamma regression model with constant shape parameter, and ‘alternate’ refers to the alternate parameterization (constant scale).

Table 1

True Model	n	Pooling Method	Selected Model		
			Lognormal	Gamma (std)	Gamma (alt)
Lognormal	650	Full Dataset	100.0	0.0	0.0
	400	Random Sample	100.0	0.0	0.0
		x-homogeneous Pools	99.9	0.1	0.0
	150	Random Sample	97.0	2.6	0.4
Gamma (standard)	650	Full Dataset	0.0	80.4	19.6
	400	Random Sample	0.0	73.5	26.5
		x-homogeneous Pools	0.0	76.8	23.2
Gamma (alternate)	150	Random Sample	1.2	60.3	38.5
	650	Full Dataset	0.0	12.5	87.5
Gamma (alternate)	400	Random Sample	0.0	19.5	80.5
		x-homogeneous Pools	0.0	17.8	82.2
	150	Random Sample	0.7	33.6	65.7
		x-heterogeneous Pools	1.9	27.4	70.7

Table 2

Simulation results comparing model chosen via AIC ('AIC Model') with the true model ('True Model'), for outcomes generated under lognormal, standard gamma, and alternate gamma regression models. "SD" is the empirical standard deviation and " $\hat{SE}$ " the mean estimated standard error.

Mean Bias ($\hat{SE}/SD$), 95% CI Coverage				
Model	β_1	β_2	β_3	β_4
<i>Lognormal</i>				
Full Data				
True Model	0.000 (1.00), 94.8	0.000 (1.00), 95.3	0.000 (1.00), 95.1	0.001 (0.99), 94.8
AIC Model	0.000 (1.00), 94.8	0.000 (1.00), 95.3	0.000 (1.00), 95.1	0.001 (0.99), 94.8
x-homogeneous Pools				
True Model	0.000 (1.00), 94.7	0.000 (1.00), 95.0	0.000 (0.99), 95.3	0.000 (0.99), 94.4
AIC Model	0.000 (1.00), 94.7	0.000 (1.00), 95.0	0.000 (0.99), 95.3	0.000 (0.99), 94.4
x-heterogeneous Pools				
True Model	0.000 (0.99), 94.5	-0.001 (0.98), 94.3	-0.003 (0.98), 94.7	0.000 (0.96), 94.0
AIC Model	0.000 (0.98), 94.2	-0.001 (0.98), 94.0	-0.004 (0.97), 94.5	0.000 (0.96), 93.8
<i>Gamma (standard)</i>				
Full Data				
True Model	0.000 (0.99), 94.9	0.000 (0.98), 94.7	-0.003 (1.00), 95.0	-0.001 (0.99), 94.8
AIC Model	0.000 (0.94), 92.8	-0.001 (0.94), 93.6	-0.007 (0.94), 92.9	0.002 (0.94), 93.5
x-homogeneous Pools				
True Model	0.000 (0.99), 94.8	0.000 (0.98), 94.7	-0.003 (1.00), 94.9	-0.001 (0.98), 94.6
AIC Model	0.000 (0.95), 93.4	-0.001 (0.95), 93.6	-0.006 (0.95), 93.4	0.001 (0.95), 93.5
x-heterogeneous Pools				
True Model	0.000 (0.98), 94.2	0.000 (0.97), 94.4	-0.003 (0.98), 94.5	-0.001 (0.97), 94.4
AIC Model	0.000 (0.95), 93.2	-0.001 (0.95), 93.7	-0.004 (0.96), 93.6	0.000 (0.96), 93.9
<i>Gamma (alternate)</i>				
Full Data				
True Model	0.000 (1.00), 95.0	0.000 (0.99), 94.6	-0.001 (1.00), 95.4	0.001 (0.98), 94.8
AIC Model	0.001 (0.97), 94.3	0.001 (0.94), 93.8	0.005 (0.99), 95.2	-0.003 (0.94), 93.7
x-homogeneous Pools				
True Model	0.000 (0.99), 94.9	0.000 (0.98), 94.3	-0.001 (1.01), 95.6	0.001 (0.98), 94.6
AIC Model	0.001 (0.95), 94.0	0.001 (0.94), 93.3	0.005 (0.99), 95.2	-0.003 (0.94), 93.6
x-heterogeneous Pools				
True Model	0.000 (0.98), 94.8	0.000 (0.97), 94.2	-0.002 (0.99), 95.1	-0.001 (0.97), 94.3
AIC Model	0.001 (0.95), 93.9	0.001 (0.94), 93.4	0.003 (0.98), 94.6	-0.003 (0.94), 93.6

Table 3

Simulation results comparing precision of estimates from the model chosen via AIC (AIC Model) with those calculated under the true model (True Model), for outcomes generated under lognormal, standard gamma, and alternate gamma regression models. Lower empirical standard deviation indicates better precision.

Pool Type	<i>n</i>	Model	β_1	β_2	β_3	β_4
<i>Lognormal</i>						
Full Data	650	True Model	0.006	0.079	0.086	0.080
		AIC Model	0.006	0.079	0.086	0.080
x-homogeneous Pools	400	True Model	0.006	0.082	0.090	0.083
		AIC Model	0.006	0.082	0.090	0.083
x-heterogeneous Pools	150	True Model	0.007	0.090	0.097	0.092
		AIC Model	0.007	0.090	0.097	0.092
<i>Gamma (standard)</i>						
Full Data	650	True Model	0.005	0.072	0.077	0.072
		AIC Model	0.006	0.072	0.078	0.072
x-homogeneous Pools	400	True Model	0.005	0.072	0.077	0.072
		AIC Model	0.005	0.072	0.078	0.072
x-heterogeneous Pools	150	True Model	0.005	0.072	0.077	0.072
		AIC Model	0.006	0.073	0.078	0.072
<i>Gamma (alternate)</i>						
Full Data	650	True Model	0.005	0.062	0.066	0.063
		AIC Model	0.005	0.067	0.069	0.068
x-homogeneous Pools	400	True Model	0.005	0.068	0.070	0.068
		AIC Model	0.005	0.073	0.074	0.073
x-heterogeneous Pools	150	True Model	0.005	0.075	0.077	0.076
		AIC Model	0.006	0.078	0.081	0.079

Table 4

Regression estimates and estimated standard errors on data from the CPP dataset, with cytokine IP as the outcome.

Fit	AIC	Age [yrs]	Smoking Status [yes/no]	Race [black/white]	SA Status [yes/no]
<i>Full Data (n=645)</i>					
Lognormal	-3879	0.009 (0.005)	0.054 (0.071)	0.179 (0.078)*	-0.045 (0.072)
Gamma (std)	-3769	0.012 (0.005)*	0.036 (0.069)	0.166 (0.075)*	-0.102 (0.069)
Gamma (alt)	-3765	0.005 (0.004)	0.037 (0.057)	0.120 (0.061)*	-0.031 (0.057)
<i>Observed Pools (n=397)</i>					
Lognormal	-2418	0.016 (0.008)	0.055 (0.114)	0.299 (0.124)*	-0.049 (0.092)
Gamma (std)	-2359	0.024 (0.009)*	0.011 (0.133)	0.347 (0.122)*	-0.105 (0.088)
Gamma (alt)	-2352	0.011 (0.006)	0.028 (0.089)	0.185 (0.095)	-0.033 (0.073)
<i>Expected Pools (n=397)</i>					
Lognormal	-2427	0.012 (0.008)	0.078 (0.104)	0.307 (0.112)*	-0.032 (0.082)
Gamma (std)	-2339	0.012 (0.009)	0.130 (0.161)	0.361 (0.118)*	-0.126 (0.082)
Gamma (alt)	-2334	0.009 (0.006)	0.045 (0.085)	0.196 (0.091)*	-0.026 (0.070)

A ‘*’ indicates significant association at the 0.05 level. ‘std’ refers to the standard parameterization of the gamma regression model, and ‘alt’ refers to the alternate parameterization.

Table 5

Regression estimates and estimated standard errors on data from the EAGeR study, with inhibin as the outcome.

Fit	AIC	Number of Previous Losses	Relative with Autoimmune Disease [yes/no]	Education [$>$ High School]	AMH [ng/mL]
<i>Full Data (n=1142)</i>					
Lognormal	11965	0.018 (0.041)	0.107 (0.058)	-0.141 (0.058)*	0.076 (0.007)*
Gamma (std)	11656	0.025 (0.034)	0.077 (0.048)	-0.067 (0.048)	0.056 (0.006)*
Gamma (alt)	11596	0.015 (0.031)	0.070 (0.043)	-0.123 (0.045)*	0.057 (0.004)*
<i>Artificially-Formed Pools (n=500)</i>					
Lognormal	4892	0.024 (0.037)	0.098 (0.052)	-0.106 (0.052)*	0.066 (0.006)*
Gamma (std)	4788	0.025 (0.033)	0.079 (0.046)	-0.067 (0.046)	0.056 (0.006)*
Gamma (alt)	4756	0.019 (0.031)	0.073 (0.042)	-0.101 (0.044)*	0.054 (0.004)*

A ** indicates significant association at the 0.05 level. 'std' refers to the standard parameterization of the gamma regression model, and 'alt' refers to the alternate parameterization.