Published in final edited form as:

Stat Med. 2009 December 30; 28(30): 3811–3828. doi:10.1002/sim.3754.

Flexible modeling of longitudinal highly-skewed outcomes

Huichao Chen¹, Amita K. Manatunga², Robert H. Lyles², Limin Peng², and Michele Marcus³

Huichao Chen: hchen@sdac.harvard.edu

¹ Center for Biostatistics in AIDS Research, Department of Biostatistics, Harvard School of Public Health, Harvard University, 651 Huntington Avenue, FXB-502, Boston, MA 02115, USA, Phone: (617) 432-0040, FAX: (617) 432-3163

² Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA

³ Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA

Abstract

The analysis of data from epidemiologic and environmental studies present challenges such as skewness of distribution, rounding and multiple measurements over time. To model trends over time based on repeated measurements, we propose a general latent model suitable for highly skewed data. The model assumes that the observed outcome is determined by an unobservable outcome which follows a Weibull distribution. To accommodate correlations among repeated responses over time, we introduce a general random effect from the power variance function (PVF) family of distributions, including the gamma distribution often employed in the literature. The resulting marginal likelihood has a closed form without resorting to numerical or approximation methods. We study estimation and hypothesis testing under these models, with different choices of random effect distributions. Simulation studies are conducted to evaluate their performance. Finally, we apply the proposed method to exposure data collected from the Michigan polybrominated biphenyl (MIPBB) study.

Keywords

grouped data; random effects; PVF; PBB; MIPBB

1 Introduction

In many environmental and epidemiologic studies, repeated exposure measurements are available on subjects and modeling the exposure over time is of interest. It is also of interest to investigate whether the change of exposure is related to subject characteristics. Motivated by a specific study known as the Michigan polybrominated biphenyl (MIPBB) study, we propose to develop a class of parametric models suitable for such longitudinal data. In the MIPBB, subjects were exposed to a chemical called polybrominated biphenyl (PBB) through an accidental contamination of cattle feed in Michigan in 1973 [1]. Since PBB is a chemical insoluble in water but soluble in fat, PBB remains in the human body for many years. The goal of the study was to examine the health effects of exposure to PBB thought to interfere with the functioning of the hormone estrogen. Since estrogen is more important among women than men, the original study was focused on women. Specifically, researchers are interested in determining the pattern of PBB decay and its association with characteristics such as body mass index (BMI). In this study women were followed for an average of 7 years, and each woman was measured from 2 to 10 times during the study period.

There are several challenges in modeling the PBB exposure over time. Figure 1 displays a frequency plot of thousands of serum PBB measurements taken longitudinally on women in the MIPBB. Prominent features of these data include right skewness (clearly present in the figure with more than 80% of PBB values concentrated between 0 parts per billion (ppb) and 10 ppb). The shaded bar on the left indicates an abundance of non-detectables (values below 1 ppb) [2]. In addition, there is clear heaping at integer values with relatively few observations occurring between integers, a characteristic of coarse data [3]. The exact data generation mechanism from the laboratory assay is unknown, although it is a reasonable assumption that the laboratory assays tend to approximate exposure measurements to the nearest integers. Due to the nature of these data, we take an approach that models the distribution of the true but generally unobserved underlying continuous exposures. This underlying true exposure distribution is assumed to generate the observed data. Due to the sparseness of values of PBB as a result of heaping due to rounding, we divide the observed PBB data into 12 intervals : < 1.0 , $1.0 - 1.5$, $1.5 - 2.5$, ..., ≥ 10.5 . As we show later, this view offers a natural approach to deal with left censoring due to non-detectables, as well as coarse data.

In the literature, when the observed data consist of several intervals, there are two main approaches for analyzing the data. One method is to analyze the midpoints of the observed intervals as if they were the true observed values. Although this approach is practically convenient, it fails to account for the grouping effect, which may produce a significant bias [4,5]. The other approach is to base inference directly on a distribution for the underlying true value and the likelihood arising from it. Compared with the first method, this approach is considered more desirable, especially for data sets which are large and coarsely grouped [6]. Previously, computational difficulties were an obstacle to implementing this approach. However, the advanced development of computational software makes routine application of this method possible [7]. Random effects models have been adopted in the literature when grouped data are correlated [8,9,10]. These methods typically involve numerical or Monte Carlo integration techniques for computing the likelihood. In this paper, we propose a latent model for modeling repeated grouped exposure measurements assuming that the observed exposure is determined by a latent variable that follows a Weibull distribution. We choose the Weibull distribution for its flexibility and because its long right tail is effective for modeling the shape of highly skewed data. Our latent model includes an unobserved random intercept which accounts for correlations among repeated responses. The distribution of the random effect is assumed to be within the power variance function (PVF) family suggested by Hougaard [11]. This family allows more flexible modeling of the random effect, and includes gamma and inverse Gaussian distributions as special cases. We obtain a closed form for the marginal likelihood without resorting to numerical or approximation methods (i.e., the marginal likelihood does not contain integration signs). This general modeling strategy also enables tests for the adequacy of the random effect distribution within the PVF family. We investigate the use of the likelihood ratio test in determining the distribution of the random effect in Section 3.2.

The method presented in this article has connections with other literature concerning ordered categorical data analysis and survival analysis. Grouped data that we refer to in this manuscript can be viewed as ordered categorical data, where categories are based on known cutpoints of a continuous scale. In the context of ordered categorical data, Crouchley [12] presented a random effects model for clustered data assuming a complementary log-log link function and used the PVF family to model the random effect. He derived the marginal distribution in a closed form for the case when only cluster-specific covariates are present. Our model can be considered as an extension of Crouchley's model when subject-specific covariates are present, and when interval cutpoints are known. In a different context, other authors developed a general expression for the closed-form likelihood under the assumption

of gamma random effects for modeling discrete event time data [13]. Also, PVF random effects distributions have been used to model correlated continuous survival data [11]. Extensions of those models to interval-censored data would provide a close resemblance to the model presented here.

The remainder of this article is organized as follows. In Section 2, we introduce the notation, formulate the model, and derive the marginal likelihood function. Estimation and hypothesis testing procedures are discussed as well. In Section 3, we evaluate the performance of the proposed models by simulation studies. Since the comparison of PVF random effects vs. gamma random effects occurs at the boundary, we also investigate the appropriateness of the one-sided likelihood ratio test via simulations in Section 3. In Section 4, we model the data from the MIPBB and compare the results under different groupings of PBB exposures. We conclude with some remarks in Section 5.

2 RANDOM EFFECTS MODEL

2.1 Notation

Suppose we have repeated measurements which are observed in groups or intervals. Let Z_{ij} denote the observed group for the j th measurement of the i th subject and X_{ij} be the corresponding vector of covariates ($i = 1, 2, \dots, I; j = 1, 2, \dots, n_i$). Suppose that Z_{ij} indicates membership in one of r intervals, where r is a finite number of partitions of the positive real line by means of known cutpoints $0 = \alpha_0 < \alpha_1 < \dots < \alpha_{r-1} < \alpha_r = \infty$. Let Y_{ij} be the underlying unobserved continuous variable associated with the grouping. The relationship between Z_{ij} and Y_{ij} is

$$\alpha_{k-1} \leq Y_{ij} < \alpha_k \iff Z_{ij} = k,$$

where $k \in \{1, \dots, r\}$. That is, the unobservable Y_{ij} belongs to the k th interval if and only if $\alpha_{k-1} \leq Y_{ij} < \alpha_k$, where α_{k-1} and α_k denote the lower and upper cutpoints of the k th interval. Since the data are highly skewed, we assume that Y_{ij} is Weibull distributed conditioning on a random effect. The random effect w_i accounts for the within-subject correlation and it has a distribution function $F(w)$ with a positive support. Conditional on w_i , it is assumed that

$$Y_{ij}|w_i \sim \text{Weibull}(w_i \eta_{ij}, \lambda), \quad (1)$$

where $\lambda > 0$ is a shape parameter and $\eta_j = \exp(-\lambda\beta_0 - \lambda\beta_1 X_{ij})$. Conditional on w_i , model (1) is equivalent to the following linear regression model:

$$\begin{aligned} \log Y_{ij} &= -\frac{1}{\lambda} \log(w_i \eta_{ij}) + \frac{1}{\lambda} e_{ij} \\ &= -\frac{1}{\lambda} \log w_i + \beta_0 + \beta_1 X_{ij} + \frac{1}{\lambda} e_{ij}, \end{aligned} \quad (2)$$

where e_{ij} has a log unit exponential distribution. This model provides an appealing interpretation of a log linear relationship between Y_{ij} and X_{ij} . The slope β_1 represents the effect of covariate X_{ij} . Each individual has a random intercept of $-\frac{1}{\lambda} \log(w_i) + \beta_0$ and the

expectation of Y_{ij} given w_i is $E(Y_{ij}|w_i) = \frac{\Gamma(1 + \frac{1}{\lambda})}{w_i^{\frac{1}{\lambda}} e^{-\beta_0 - \beta_1 X_{ij}}}$.

2.2 Marginal Likelihood

Based on the assumption that responses from the same subject are independent conditional on the random effect w_i , we express the likelihood function for the i th individual from model (2) as follows:

$$\begin{aligned} L_i^*(\beta_0, \beta_1, \lambda | w_i) &= \prod_{j=1}^{n_i} \prod_{k=1}^r \Pr(Z_{ij}=k | w_i)^{\delta_{ijk}} \\ &= \prod_{j=1}^{n_i} \prod_{k=1}^r \Pr(\alpha_{k-1} \leq Y_{ij} < \alpha_k | w_i)^{\delta_{ijk}} \\ &= \prod_{j=1}^{n_i} \prod_{k=1}^r (\pi_{i,j,k-1} - \pi_{ijk})^{\delta_{ijk}}, \end{aligned} \quad (3)$$

where $\pi_{ijk} = \Pr(Y_{ij} \geq \alpha_k | w_i) = e^{-w_i \eta_{ij} \alpha_k^\lambda}$, $\eta_{ij} = \exp(-\lambda \beta_0 - \lambda \beta_1 X_{ij})$, $\pi_{ij0} = 1$, $\pi_{ijr} = 0$, and δ_{ijk} is an index variable which equals 1 when $Z_{ij} = k$, otherwise 0. Although we assume that there is an underlying continuous Weibull distribution, the only assumption needed is $\pi_{ijk} = e^{-w_i \eta_{ij} \alpha_k^\lambda}$. In other words, we assume that conditional on w_i ,

$$\log(-\log \pi_{ijk}) = \log w_i - \lambda \beta_0 - \lambda \beta_1 X_{ij} + \lambda \log \alpha_k, \quad (4)$$

where the α_k 's are known constants. By mathematical derivations shown in Appendix 7.1, the conditional likelihood for subject i takes the form

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i, r-1}=0}^{\delta_{in_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} e^{-w_i S_i},$$

where $S_i = \sum_{j=1}^{n_i} \sum_{k=1}^{r-1} D_{ijk}$ and

$$\begin{aligned} D_{ijk} &= \eta_{ij} \cdot \alpha_k^\lambda (\delta_{ijk} - b_{ijk} + b_{i,j,k+1}) \quad \text{for } k \leq r-2 \\ D_{ijk} &= \eta_{ij} \cdot \alpha_{r-1}^\lambda (\delta_{i,j,r-1} - b_{i,j,r-1} + \delta_{ijr}) \quad \text{for } k = r-1. \end{aligned}$$

It follows that the marginal likelihood for the i th subject is

$$L_i(\beta_0, \beta_1, \lambda) = \int L_i^*(\beta_0, \beta_1, \lambda | w_i) dF(w_i),$$

which can be written in terms of the expectation of $e^{-w_i S_i}$ as

$$L_i(\beta_0, \beta_1, \lambda) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i, r-1}=0}^{\delta_{in_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} E(e^{-w_i S_i}).$$

Since the random effect w_i has a positive support, the above likelihood equation can be explicitly written when $E(e^{-w_i S_i}) = p(S_i)$, the Laplace transform of w_i . A common choice for the random effect distribution is log-normal but in this case the Laplace transform cannot be explicitly written. When w_i is log-normal distributed, Gaussian quadrature can be used to integrate out the random effects as illustrated by [14] for an ordered probit model. When w_i is assumed to be gamma distributed with mean 1 and variance $\frac{1}{\psi}$, say $w_i \sim \text{Gamma}(\psi, \psi)$,

then the Laplace transform is $E(e^{-S_i w_i}) = \left(\frac{\psi}{\psi + S_i}\right)^\psi$, and the likelihood takes the form

$$L_i(\beta_0, \beta_1, \lambda, \psi) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i, r-1}=0}^{\delta_{in_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} \left(\frac{\psi}{\psi + S_i} \right)^\psi. \quad (5)$$

Ten Have [13] assumed gamma distributed random effects and obtained a closed-form marginal likelihood for discrete time survival data. Hougaard [15] introduced a more general family of distributions called the PVF family. When w_i follows the PVF distribution with mean 1 and variance $\frac{1-\gamma}{\psi}$, we denote $w_i \sim \text{PVF}(\gamma, \psi^{1-\gamma}, \psi)$. The density function of w_i is as follows:

$$f(w_i) = -\frac{1}{\pi w_i} \exp\left(-\psi w_i + \frac{\psi}{\gamma}\right) \sum_{k=1}^{\infty} \frac{\Gamma(k\gamma+1)}{k!} \left(-\frac{\psi^{1-\gamma} w_i^{-\gamma}}{\gamma}\right)^k \sin(\gamma k \pi),$$

where $0 < \gamma \leq 1$ and $\psi > 0$. The Laplace transform is

$$E(e^{-S_i w_i}) = \exp\left[-\frac{\psi^{1-\gamma}}{\gamma} \{(\psi + S_i)^\gamma - \psi^\gamma\}\right].$$

Therefore, the likelihood can be expressed as

$$L_i(\beta_0, \beta_1, \lambda, \psi, \gamma) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i, r-1}=0}^{\delta_{in_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} \exp\left[-\frac{\psi^{1-\gamma}}{\gamma} \{(\psi + S_i)^\gamma - \psi^\gamma\}\right]. \quad (6)$$

The PVF distribution family includes the gamma distribution as a special case when $\gamma \rightarrow 0$, and the inverse Gaussian distribution when $\gamma = \frac{1}{2}$ (see [15]). In this article, we consider the PVF distribution for modeling repeated exposure data because of its flexibility. Another advantage of our approach is that the likelihood can be explicitly written. Therefore, likelihood ratio tests and confidence intervals for the parameters can be obtained without numerical approximations.

2.3 Prediction

In many environmental health exposure studies with multiple measurements over time, it is of interest to predict subjects' exposure at a later time point. For example, MIPBB researchers are interested in predicting the in-utero serum PBB levels of exposed women which may influence the health outcomes of their children. Under the proposed modeling

approach, we can conveniently estimate the random effect w_i for a subject given that subject's data. The predicted random effect is obtained via the posterior mean as follows:

$$\begin{aligned}\widehat{w}_i &= E(w_i | Z_i = z_i) \\ &= \int w_i f(w_i | z_i) dw_i \\ &= \frac{\int w_i f(z_i | w_i) f(w_i) dw_i}{f(z_i)}\end{aligned}$$

As $f(z_i | w_i)$ is equal to the conditional likelihood L_i^* as shown in Section 2.2, we have

$$\begin{aligned}\widehat{w}_i &= \frac{\int w_i L_i^* f(w_i) dw_i}{L_i} \\ &= \frac{\int \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[\left(\prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \right) \cdot w_i e^{-S_i \cdot w_i} f(w_i) \right] dw_i}{\sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[\left(\prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \right) \cdot \Phi^*(S_i, \psi, \gamma) \right]} \\ &= \frac{L_i}{L_i}\end{aligned}$$

where

$$\Phi^*(S_i, \psi, \gamma) = E(w_i \cdot e^{-S_i \cdot w_i}) = \begin{cases} \left(\frac{\psi}{\psi + S_i} \right)^{\psi+1}, & \text{if } \gamma = 0 \\ \exp \left[-\frac{\psi^{1-\gamma}}{\gamma} ((\psi + S_i)^\gamma - \psi^\gamma) \right] \cdot \left(\frac{\psi}{\psi + S_i} \right)^{1-\gamma}, & \text{if } \gamma \in (0, 1]. \end{cases}$$

2.4 Estimation and Hypothesis Testing

The unknown parameters $\theta = (\beta_0, \beta_1, \lambda, \psi, \gamma)$ can be estimated by the method of maximum likelihood, where $\lambda > 0$, $\psi > 0$, and $0 < \gamma \leq 1$. Maximum likelihood estimators (MLEs) have some desirable properties. Under certain regularity conditions, MLEs found by solving a set of score equations $\frac{\partial \log L}{\partial \theta} = 0$ have an asymptotically normal distribution with means θ and

covariance matrix $\{-E(\frac{\partial^2 \log L}{\partial \theta \partial \theta'})\}^{-1}$. Although the marginal likelihood functions (5) and (6) shown in the previous section are explicit, score equations do not provide explicit expressions for MLEs. Thus, we use the quasi-Newton method for maximization. When the random effect is gamma distributed, the components of the score and Hessian matrix corresponding to the likelihood are presented in Appendix 7.2. Similar expressions can be derived for other random effect distributions. Based on the formula in Section 2.3, the random effect w_i can be estimated given MLEs of θ .

An advantage of the PVF random effects model is that it includes the gamma distribution as a special case, $\gamma \rightarrow 0$. Therefore, we consider a hypothesis test for $H_0: \gamma = 0$ versus $H_a: \gamma > 0$. Under the null hypothesis, the true parameter value is on the boundary of the parameter space, which means the customary asymptotic χ_1^2 distribution for the likelihood ratio test may not hold. Previous work by [16] presented a variety of examples where the limiting distributions of likelihood ratio test statistics are mixtures of chi-squared distributions. Some authors in [17] discussed the problem of testing for nonzero variance components in the longitudinal mixed effects model of [18] and concluded that a mixed chi-square asymptotic

distribution is appropriate. For comparing the gamma vs. PVF distribution (testing $H_0: \gamma = 0$), the appropriate null distribution for model (6) is potentially a mixed chi-square as in other cases. We investigate the behavior of one-sided likelihood ratio test for the PVF distribution with simulation studies in Section 3.2.

3 SIMULATIONS

3.1 Parameter Estimation

We conduct a simulation study to evaluate the performance of the proposed model (6). Five hundred simulations are performed for two sample sizes ($I = 500, 1000$). In each simulation study, the latent variable Y_{ij} is generated from a Weibull($w_i \eta_{ij}, \lambda$) distribution with subject-specific random effect $w_i \sim \text{PVF}(\gamma, \psi^{1-\gamma}, \psi)$. Here $\eta_{ij} = e^{-\lambda \beta_0 - \lambda \beta_1 X_{ij}}$, $i = 1, \dots, I, j = 1, \dots, n_i$, $(\beta_0, \beta_1, \lambda, \psi, \gamma) = (0.05, -0.01, 5.0, 0.1, 0.5)$. The number of measurements n_i is set as 4. The covariate X_{ij} corresponding to the four measurements are set to be 0, 1, 2, 3. The data are then grouped into the following intervals: $< 1.0, 1.0 - 1.5, 1.5 - 2.5, \dots, \geq 10.5$.

Table 1 presents the average of the following quantities from 500 simulations for each parameter: the maximum likelihood parameter estimate, the estimated standard error (SE), the empirical standard deviation (SD), and the coverage probability of the nominal 95% CI (Cov95). In all cases, the parameter bias is small, decreasing as I increases. The estimated standard error and the empirical standard deviation agree well. The coverage probabilities are generally satisfactory and perform well with moderate sample sizes.

3.2 Evaluation of the Likelihood Ratio Test in PVF Distribution

Simulation studies are carried out to examine the performance of the likelihood ratio test for comparing the gamma model (5) and the PVF model (6), specifically testing the hypothesis

$$H_0: \gamma=0 \quad \text{versus} \quad H_a: \gamma>0.$$

Both χ_1^2 distribution and the mixed chi-square distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ are assessed, respectively, as the distribution of the test statistic. A set of 500 simulations is performed for different sample sizes ($I = 100, 500, 1000, 2000$). For each of the simulations, we generate Y_{ij} from the same distribution, namely, Weibull($w_i \eta_{ij}, \lambda$) with subject-specific random effect $w_i \sim \text{PVF}(\gamma, \psi^{1-\gamma}, \psi)$. Here $\eta_{ij} = e^{-\lambda \beta_0 - \lambda \beta_1 X_{ij}}$, $i = 1, \dots, I, j = 1, \dots, n_i$, $n_i = 4$, $(\beta_0, \beta_1, \lambda, \psi, \gamma) = (0.05, -0.01, 5.0, 0.1, 0)$, and $X_{i1} = 0, X_{i2} = 1, X_{i3} = 2$, and $X_{i4} = 3$ for all i . The data are then grouped into 12 intervals.

Table 2 summarizes the simulation results. In order to describe the bias and the spread of the estimators of γ , based on the distribution of estimated gamma (500 values), the mean, 5% and 95% quantiles are calculated. Since the true value of γ is zero, the mean indicates the bias and quantiles indicate the spread of the estimated distribution. Sizes of the likelihood ratio tests are also estimated.

The bias and 90% interquantile range decrease steadily as the sample size increases. In all cases, the size of the likelihood ratio test from the χ_1^2 distribution stays well below 5%, while that from the mixed chi-square is closer to 5%. Therefore, the χ_1^2 based likelihood ratio test seems conservative. This is due to the fact that, under null hypothesis, the value for γ is at the boundary of the parameter space. Thus, the approximation of the log likelihood ratio statistic by a single χ_1^2 distribution is inadequate, and it appears more reasonable to use

a mixed chi-square distribution. Figure 2 displays the estimated density function of the log likelihood ratio when the sample size equals 1000.

Results from other sample sizes are similar and therefore omitted. It appears that from Figure 2, a 50:50 mixture of χ_0^2 and χ_1^2 distribution for the likelihood ratio test is reasonable in model selection between the PVF and the gamma distributions.

4 APPLICATION TO THE MICHIGAN MIPBB STUDY

We use the proposed models to analyze PBB exposure measurements from the MIPBB. This analysis included a total of 2687 observations contributed by 955 women over the year 1976 to 1993. Each woman was measured from 2 to 10 times, with a median of 2 times. The mean PBB level at enrollment was 11.2 ppb, which is much higher than that for the general U.S. population (less than 1.0 ppb). The observed PBB levels ranged from less than 1.0 to 1114.2 with a median of 2.0, indicating the high skewness of the distribution. Of all the observations, around 20% were left-censored due to the laboratory detection limit of 1.0 ppb, and around 70% were heaped data, concentrated mainly at integers from 1 to 10. Others were continuous coarse data sparsely distributed in between integers.

Figure 1 demonstrates the histogram of the PBB measurements in the log transformed scale, $\log(\text{PBB}+1)$. The distribution is skewed after the log-transformation and an underlying Weibull distribution appears to be a reasonable approximation. As described in the introduction, PBB values are concentrated around integers etc., therefore we group the data into following intervals: < 1.0 , $1.0-1.5$, $1.5-2.5$, ..., ≥ 10.5 . Let Y_{ij} denote $\log(\text{PBB}_{ij}+1)$, where PBB_{ij} is the j th PBB measurement on the i th woman in equation (2). We define t_{ij} as the time in years since her first measurement. We consider a single covariate the body mass index (BMI) measured at the first visit (baseline). The average baseline BMI is 25 (minimum=15.0, maximum=55.5, standard deviation =4.8). For easier interpretation, we center the BMI at the mean value of 25 kg/m^2 . Assuming Y_{ij} follows a Weibull distribution conditional on the random effect w_i , we write the model as follows:

$$\log Y_{ij} = -\frac{1}{\lambda} \log w_i + \beta_0 + \beta_1 t_{ij} + \beta_2 (\text{BMI}_i - 25) + \beta_3 t_{ij} \times (\text{BMI}_i - 25) + \frac{1}{\lambda} e_{ij}.$$

It is important to note that the transformation of the data is not needed for the estimation of the parameters from model (2), as only the cutpoints are affected by the transformation. Assuming that w_i follows a PVF distribution, we maximize the likelihood given in (6). Since the PVF random effect model reduces to gamma random effect model when $\gamma = 0$, we fit both models to the data. The results are summarized in Table 3.

Based on the Wald test for the null hypothesis $H_0 : \gamma = 0$, we conclude that the model with PVF random effects fits significantly better than the model with gamma random effects ($Z=6.18$, $p < 0.0001$). This is also supported by the classical likelihood ratio test with chi-squared distribution ($\text{LR}=15.70$, $p < 0.0001$). As indicated in the Section 3.2, we also use a mixed chi-square reference distribution ($\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$), which does not alter the significance of the result. Surprisingly, regression coefficients and standard errors are very similar from both models. The estimated slope in the PVF model is -0.002 ($p < 0.05$) from a woman with average BMI of 25 kg/m^2 . Direct interpretation of these parameters are somewhat difficult for the latent model (2) since we have a log transformation, but can be done via the expectation of the Y_{ij} (given below the model (2)). However, the interpretation for the model (4) is straightforward. That is, the slope for this model becomes 0.012. This implies that for a given subject, the probability of falling above a certain cut point of exposure after one year

will decrease to the power of $1.012(\exp(0.012)=1.012)$ of the corresponding probability for the initial exposure. Using the expectation given for the latent model (2), the average PBB (in log-scale) for a given subject with an average BMI of 25 kg/m^2 after one year will be $\exp(-0.002) = 0.998$ times of that woman's initial PBB values. From Table 3, the estimated β_3 is 0.001 ($p < 0.0001$), showing different time trends for women with different initial BMI measurements. Women with higher values of BMI tend to have a slower decrease in PBB levels compared to women with low BMI. In fact, if the model is true, there is a positive time trend for some women with large BMI values which is in agreement with the data observed. For example, there are about 30% of subjects with $\text{BMI} > 27$ in the data. For this subset of data the mean slope was -0.011 (range $-0.542, 0.835$) using the least squared methods. Our conclusions are consistent with the findings from [2] where authors analyzed the MIPBB using the least squares method, suggesting a greater retention of PBB in women with greater amounts of adipose tissue.

To assess the adequacy of the model, we calculate the observed proportions and the averages of model-based probabilities $Pr(Z_{ij} = k)$ at baseline at each category, where $i = 1, \dots, 955$; $j = 1$; and k ranges from 1 to 12. The observed and predicted proportions are (0.225, 0.173, 0.177, 0.117, 0.055, 0.043, 0.036, 0.026, 0.017, 0.019, 0.009, 0.103) and (0.257, 0.177, 0.182, 0.090, 0.054, 0.036, 0.026, 0.019, 0.015, 0.012, 0.010, 0.122), respectively. A plot of this data (not included) indicates that the average of PVF-based probabilities seems to be close to the observed proportion at the baseline.

Since the random effects w_i can be estimated for each subject i ($i = 1, 2, \dots, 955$) based on formulas in Section 2.3, we present in Figure 3 the histogram of empirical Bayes estimates (log-scale), assuming the PVF random effects model is true. To compare these values to the densities of random effects, we plot the densities of $\log(w_i)$ for PVF and Gamma distributions with estimated parameters from Table 3. This plot shows that empirical Bayes estimates from the data are closer to the PVF distribution than the gamma distribution. In the context of linear mixed model, Verbeke and Molenberghs [20](page 84) points out that caution should be exercised as empirical Bayes estimates are highly dependent on the assumed prior distribution.

We also conduct analyses under different groupings of PBB exposures to check the sensitivity to grouping. Some of the results are as follows. When we group the data into 11 intervals: $< 1.0, 1.0-1.5, 1.5-2.5, \dots, \geq 9.5$, the estimated β_1 is -0.002 ($p < 0.0001$) and the estimated β_3 is 0.001 ($p < 0.0001$). If we group the data into 13 categories: $< 1.0, 1.0-2.0, 2.0-3.0, \dots, \geq 12$, the estimates for β_1 and β_3 are -0.001 ($p < 0.0001$) and 0.001 ($p < 0.0001$), respectively. Compared to results in Table 3, we observe that, under all of these different groupings, the PBB level significantly decreases over time ($\beta_1 = -0.002, -0.002, -0.001$) and the interaction between time and BMI is significant too ($\beta_3 = 0.001, 0.001, 0.001$). These results under different groupings indicate that our approach of grouping PBB exposure data is practically reliable.

5 DISCUSSION

We have presented a random effects model for the analysis of highly skewed grouped longitudinal data. Development of the model assumes that the observed groups are generated by an underlying Weibull distribution which leads to a complementary log-log link for the probability of the outcome exceeding an ordinal cutpoint. However, the Weibull assumption is not critical in estimation of parameters. In fact, in order to derive the likelihood, we only need to assume that a complementary log-linear model holds conditional on the random effect. However, if one is interested in estimating the initial exposure or to predict an exposure at a given time in the continuous scale, the Weibull assumption is needed.

We have used a flexible PVF family random effects distribution to account for correlations among repeated measurements. Under this PVF random effects model with a complementary log link, we obtain closed-form likelihoods after taking into account grouped nature of the data. If the underlying structure is assumed to be normally distributed conditional on normal random effects, the corresponding model for the grouped data is a probit-link model with known cut points. Under this model with normal random effects the marginal likelihood function cannot be explicitly written. To obtain maximum likelihood estimates, numerical integration techniques are required. Since both models (Weibull and Normal) operate under a certain set of assumptions, they are expected to perform differently. While it could be interesting to compare such models for a given dataset, the issue of goodness of fit across different parametric families is beyond the scope of this manuscript.

Another significant advantage of our approach is the closed form of the marginal likelihoods, allowing convenient maximum likelihood estimation of parameters and their standard errors. Several authors [14,19,21] have considered modeling ordered categorical data with repeated measurements by considering a normal random effects distribution. Others [22,23,24] have used a non-parametric distribution for random effects, thereby providing more robust semi-parametric models. In many of these cases, closed-form likelihoods are not available and approximations must be used. The model presented here can be considered as an alternative for analyzing such data, offering flexibility for characterizing skewness along with the advantage of a closed-form likelihood.

Our method operates under parametric assumptions where the underlying latent distribution is assumed to be Weibull distributed. Since the interval cut-points are observed and known, data can help to confirm the shape of this distribution. However, there is not much information in the right tail of the distribution beyond uppermost interval cutpoint. Therefore the tail behavior is effectively dictated by the parametric assumption. This consequence is commonly encountered in parametric survival data analysis with right-censored data.

In our data analysis, according to Figure 1 we assume that the PBB values are centered around integers. Unfortunately we do not have any specific information that can explain the behavior of the data. As a reviewer points out, there may be a loss of information from rounding off the data. If the underlying mechanism of the data generation is known, it is possible to analyze the evidently rounded-off observations as though they are rounded off and the non-rounded observations as though they are exact. Such an approach can minimize the loss of information.

Acknowledgments

This research was supported by Awards R01-ES012458 and R01-ES012014 from the National Institute of Environmental Health Sciences, and was supported in part by PHS Grant, UL1 RR025008 from the Clinical and Translational Science Award program, National Institutes of Health. We thank the staff of the Michigan Department of Community Health and participants of the Michigan Long-Term PBB Study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

References

1. Blanck HM, Marcus M, Hertzberg VS, Tolbert PE, Rubin C, Henderson AK, Zhang RH. Determinants of polybrominated biphenyl serum decay among women in the michigan PBB cohort. *Environmental Health Perspectives*. 2000; 108:147–152. [PubMed: 10656855]
2. Blanck HM, Marcus M, Tolbert PE, Rubin C, Henderson AK, Hertzberg VS, Zhang RH, Cameron L. Age at menarche and tanner stage in girls exposed in utero and postnatally to polybrominated biphenyl. *Epidemiology*. 2000; 11:641–647. [PubMed: 11055623]

3. Heitjan DF, Rubin DB. Ignorability and coarse data. *The Annals of Statistics*. 1991; 19:2244–2252.
4. Hasselblad V, Stead AG, Galke W. Analysis of coarsely grouped data from the lognormal distribution. *Journal of the American Statistical Association*. 1980; 75:771–778.
5. Beckman RJ, Johnson ME. Fitting the Student t distribution to grouped data, with application to a particle scattering experiment. *Technometrics*. 1987; 29:17–22.
6. Heitjan DF. Grouped data. *Encyclopedia of Biostatistics*. 1998; 2:1784–1785.
7. Heitjan DF. Regression with bivariate grouped data. *Biometrics*. 1991; 47:549–562. [PubMed: 1912261]
8. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*. 2000; 9:161–179. [PubMed: 10946432]
9. Rabe-Hesketh S, Yang S, Pickles A. Multilevel models for censored and latent responses. *Statistical Methods in Medical Research*. 2001; 10:409–427. [PubMed: 11763550]
10. Grilli L. The random-effects proportional hazards model with grouped survival data: a comparison between the grouped continuous and continuation ratio versions. *Journal of the Royal Statistical Society, Series A*. 2005; 168:83–94.
11. Hougaard P. *Analysis of multivariate survival data*. New York: Springer; 2000.
12. Crouchley R. A random-effects model for ordered categorical data. *Journal of the American Statistical Association*. 1995; 90:489–498.
13. Ten Have TR. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics*. 1996; 52:473–491. [PubMed: 8672699]
14. Jansen J. On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics*. 1990; 39:75–84.
15. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*. 1986; 73:387–396.
16. Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*. 1987; 82:605–610.
17. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects models. *Biometrics*. 1994; 50:1171–1177. [PubMed: 7786999]
18. Laird N, Ware JH. Random-effects models for longitudinal data. *The American Statistician*. 1982; 38:963–974.
19. Harville DA, Mee RW. A mixed-model procedure for analyzing ordered categorical data. *Biometrics*. 1984; 40:393–408.
20. Verbeke, G.; Molenberghs, G. *Linear mixed models for longitudinal data*. New York: Springer; 2000.
21. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 1994; 50:933–944. [PubMed: 7787006]
22. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*. 1978; 73:805–811.
23. Follmann DA, Lambert D. Generalizing logistics regression by nonparametric mixing. *Journal of the American Statistical Association*. 1989; 84:295–300.
24. Tao H, Palta M, Yandell BS, Newton MA. An estimation method for the semiparametric mixed effects model. *Biometrics*. 1999; 55:102–110. [PubMed: 11318143]

7 APPENDIX

7.1 Conditional Likelihood for Subject i

Using the binomial expansion, we write the equation (3) as

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \prod_{j=1}^{n_i} \left\{ \prod_{k=1}^{r-1} \sum_{b_{ijk}=0}^{\delta_{ijk}} (-1)^{\delta_{ijk}-b_{ijk}} (\pi_{i,j,k-1})^{b_{ijk}} (\pi_{ijk})^{\delta_{ijk}-b_{ijk}} \right\} (\pi_{i,j,r-1})^{\delta_{ijr}}. \quad (7)$$

The expression within the braces in (7) can be rewritten as $(r-1)$ summations. That is,

$$\sum_{b_{ij1}=0}^{\delta_{ij1}} \cdots \sum_{b_{i,j,r-1}=0}^{\delta_{i,j,r-1}} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} (\pi_{i,j,k-1})^{b_{ijk}} (\pi_{ijk})^{\delta_{ijk}-b_{ijk}}. \quad (8)$$

After substituting (8) into (7), we change the sequence of the inner summations with the outer-most product, which gives the following:

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{i n_i, r-1}=0}^{\delta_{i n_i, r-1}} \prod_{j=1}^{n_i} \left\{ \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} (\pi_{i,j,k-1})^{b_{ijk}} (\pi_{ijk})^{\delta_{ijk}-b_{ijk}} \right\} (\pi_{i,j,r-1})^{\delta_{ijr}}.$$

We rewrite this expression as

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{i n_i, r-1}=0}^{\delta_{i n_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \right\} \times \prod_{j=1}^{n_i} \left\{ \prod_{k=1}^{r-1} (\pi_{i,j,k-1})^{b_{ijk}} (\pi_{ijk})^{\delta_{ijk}-b_{ijk}} \right\} (\pi_{i,j,r-1})^{\delta_{ijr}},$$

which is equivalent to

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{i n_i, r-1}=0}^{\delta_{i n_i, r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \right\} \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} C_{ijk},$$

where

$$C_{ijk} = \pi_{ijk}^{\delta_{ijk}-b_{ijk}+b_{i,j,k+1}} \quad \text{for } k \leq r-2$$

$$C_{ijk} = \pi_{i,j,r-1}^{\delta_{i,j,r-1}-b_{i,j,r-1}+\delta_{ijr}} \quad \text{for } k=r-1.$$

Following [13], we write

$$D_{ijk} = \eta_{ij} \cdot \alpha_k^\lambda (\delta_{ijk} - b_{ijk} + b_{i,j,k+1}) \quad \text{for } k \leq r-2$$

$$D_{ijk} = \eta_{ij} \cdot \alpha_{r-1}^\lambda (\delta_{i,j,r-1} - b_{i,j,r-1} + \delta_{ijr}) \quad \text{for } k=r-1.$$

Notice that $\pi_{ijk} = \Pr(Y_{ij} \geq \alpha_k | w_i) = e^{-w_i \eta_{ij} \alpha_k^\lambda}$, it follows that

$$C_{ijk} = e^{-w_i \eta_{ij} \alpha_k^1 (\delta_{ijk} - b_{ijk} + b_{ijk+1})} = e^{-w_i D_{ijk}} \quad \text{for } k \leq r-2$$

$$C_{ijk} = e^{-w_i \eta_{ij} \alpha_{r-1}^1 (\delta_{ij,r-1} - b_{ij,r-1} + b_{ijr})} = e^{-w_i D_{ij,r-1}} \quad \text{for } k=r-1.$$

Finally, the conditional likelihood for subject i takes the form

$$L_i^*(\beta_0, \beta_1, \lambda | w_i) = \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} e^{-w_i D_{ijk}}$$

$$= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk} - b_{ijk}} \right\} e^{-w_i S_i},$$

$$\text{where } S_i = \sum_{j=1}^{n_i} \sum_{k=1}^{r-1} D_{ijk}.$$

7.2 Score and Hessian Equations

Let $\lambda = e^{\lambda_1}$ and $\psi = e^{\psi_1}$. The score and the Hessian equations with respect to $\theta = (\beta_0, \beta_1, \lambda_1, \psi_1)$ for the marginal likelihood in (5) are

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^I \frac{1}{L_i} \frac{\partial L_i}{\partial \theta}$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \theta'} = \sum_{i=1}^I \left(\frac{1}{L_i} \frac{\partial^2 L_i}{\partial \theta \partial \theta'} - \frac{1}{L_i^2} \frac{\partial L_i}{\partial \theta} \frac{\partial L_i}{\partial \theta'} \right)$$

The terms related to the i^{th} individual can be expressed respectively as follows:

$$\begin{aligned}
\frac{\partial L_i}{\partial \beta_0} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[\left\{ \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \right\} \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \beta_0} \right] \\
&= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \cdots \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left(P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \beta_0} \right) \\
P_i &= \prod_{j=1}^{n_i} \prod_{k=1}^{r-1} (-1)^{\delta_{ijk}-b_{ijk}} \\
\frac{\partial \Phi_i}{\partial S_i} &= - \left(\frac{\psi}{\psi+S_i} \right)^{\psi+1} \\
\frac{\partial S_i}{\partial \beta_0} &= -\lambda S_i \\
\frac{\partial L_i}{\partial \beta_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left(P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \beta_1} \right) \\
\frac{\partial S_i}{\partial \beta_1} &= \sum_{j=1}^{n_i} \left(\sum_{k=1}^{r-2} D_{ijk} + D_{ij,r-1} \right) (-\lambda X_{ij}) \\
\frac{\partial L_i}{\partial \lambda_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left(P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \lambda_1} \right) \\
\frac{\partial S_i}{\partial \lambda_1} &= \sum_{j=1}^{n_i} \left[\sum_{k=1}^{r-2} D_{ijk} \{ \log(\alpha_k) - \beta_0 - \beta_1 X_{ij} \} + D_{ij,r-1} \{ \log(\alpha_{r-1}) - \beta_0 - \beta_1 X_{ij} \} \right] e^{\lambda_1} \\
\frac{\partial L_i}{\partial \psi_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \left\{ \frac{\psi}{\psi+S_i} \right\}^{\psi} \left\{ \log \left(\frac{\psi}{\psi+S_i} \right) + \frac{S_i}{\psi+S_i} \right\} \right] e^{\psi_1} \\
\frac{\partial^2 L_i}{\partial \beta_0^2} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \left\{ \frac{\partial^2 \Phi_i}{\partial S_i^2} \left(\frac{\partial S_i}{\partial \beta_0} \right)^2 + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \beta_0^2} \right\} \right] \\
\frac{\partial^2 \Phi_i}{\partial S_i^2} &= \psi^{\psi+1} (\psi+1) (\psi+S_i)^{-\psi-2} \\
\frac{\partial^2 S_i}{\partial \beta_0^2} &= -\lambda \frac{\partial S_i}{\partial \beta_0} \\
\frac{\partial^2 L_i}{\partial \beta_0 \partial \beta_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ P_i \left(\frac{\partial^2 \Phi_i}{\partial S_i^2} \frac{\partial S_i}{\partial \beta_0} \frac{\partial S_i}{\partial \beta_1} + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \beta_0 \partial \beta_1} \right) \right\} \\
\frac{\partial^2 S_i}{\partial \beta_0 \partial \beta_1} &= -\lambda \frac{\partial S_i}{\partial \beta_1} \\
\frac{\partial^2 L_i}{\partial \beta_0 \partial \lambda_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ P_i \left(\frac{\partial^2 \Phi_i}{\partial S_i^2} \frac{\partial S_i}{\partial \lambda_1} \frac{\partial S_i}{\partial \beta_0} + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \beta_0 \partial \lambda_1} \right) \right\} \\
\frac{\partial^2 S_i}{\partial \beta_0 \partial \lambda_1} &= -\lambda \frac{\partial S_i}{\partial \lambda_1} + \frac{\partial S_i}{\partial \beta_0} \\
\frac{\partial^2 L_i}{\partial \beta_0 \partial \psi_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \beta_0} \left\{ \log \left(\frac{\psi}{\psi+S_i} \right) + 1 + \frac{1}{\psi} - \frac{\psi+1}{\psi+S_i} \right\} \right] e^{\psi_1} \\
\frac{\partial^2 L_i}{\partial \beta_1^2} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \left\{ \frac{\partial^2 \Phi_i}{\partial S_i^2} \left(\frac{\partial S_i}{\partial \beta_1} \right)^2 + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \beta_1^2} \right\} \right] \\
\frac{\partial^2 S_i}{\partial \beta_1^2} &= \sum_{j=1}^{n_i} \left(\sum_{k=1}^{r-2} D_{ijk} + D_{ij,r-1} \right) \lambda^2 X_{ij}^2 \\
\frac{\partial^2 L_i}{\partial \beta_1 \partial \lambda_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ P_i \left(\frac{\partial^2 \Phi_i}{\partial S_i^2} \frac{\partial S_i}{\partial \lambda_1} \frac{\partial S_i}{\partial \beta_1} + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \beta_1 \partial \lambda_1} \right) \right\} \\
\frac{\partial^2 S_i}{\partial \beta_1 \partial \lambda_1} &= \sum_{j=1}^{n_i} \left[\left\{ \sum_{k=1}^{r-2} D_{ijk} (\log \alpha_k - \beta_0 - \beta_1 X_{ij}) + D_{ij,r-1} (\log \alpha_{r-1} - \beta_0 - \beta_1 X_{ij}) \right\} - X_{ij} \left(\sum_{k=1}^{r-2} D_{ijk} + D_{ij,r-1} \right) \right] e^{\lambda_1} \\
\frac{\partial^2 L_i}{\partial \beta_1 \partial \psi_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \beta_1} \left\{ \log \left(\frac{\psi}{\psi+S_i} \right) + 1 + \frac{1}{\psi} - \frac{\psi+1}{\psi+S_i} \right\} \right] e^{\psi_1} \\
\frac{\partial^2 L_i}{\partial \lambda_1^2} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \left\{ \frac{\partial^2 \Phi_i}{\partial S_i^2} \left(\frac{\partial S_i}{\partial \lambda_1} \right)^2 + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \lambda_1^2} \right\} \right] \\
\frac{\partial^2 S_i}{\partial \lambda_1^2} &= \sum_{j=1}^{n_i} \left[\left\{ \sum_{k=1}^{r-2} D_{ijk} (\log \alpha_k - \beta_0 - \beta_1 X_{ij})^2 + D_{ij,r-1} (\log \alpha_{r-1} - \beta_0 - \beta_1 X_{ij})^2 \right\} e^{\lambda_1} + \sum_{k=1}^{r-2} D_{ijk} (\log \alpha_k - \beta_0 - \beta_1 X_{ij}) + D_{ij,r-1} (\log \alpha_{r-1} - \beta_0 - \beta_1 X_{ij}) \right] e^{\lambda_1} \\
\frac{\partial^2 L_i}{\partial \lambda_1 \partial \psi_1} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \frac{\partial \Phi_i}{\partial S_i} \frac{\partial S_i}{\partial \lambda_1} \left\{ \log \left(\frac{\psi}{\psi+S_i} \right) + 1 + \frac{1}{\psi} - \frac{\psi+1}{\psi+S_i} \right\} \right] e^{\psi_1} \\
\frac{\partial^2 L_i}{\partial \lambda_1^2} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left[P_i \left\{ \frac{\partial^2 \Phi_i}{\partial S_i^2} \left(\frac{\partial S_i}{\partial \lambda_1} \right)^2 + \frac{\partial \Phi_i}{\partial S_i} \frac{\partial^2 S_i}{\partial \lambda_1^2} \right\} \right] \\
\frac{\partial^2 L_i}{\partial \lambda_1^2} &= \sum_{b_{i11}=0}^{\delta_{i11}} \sum_{b_{i12}=0}^{\delta_{i12}} \sum_{b_{in_i,r-1}=0}^{\delta_{in_i,r-1}} \left\{ P_i \left\{ \left(\frac{\psi}{\psi+S_i} \right) \left(\log \left(\frac{\psi}{\psi+S_i} \right) + \frac{S_i}{\psi+S_i} \right) + \left(\frac{\psi}{\psi+S_i} \right) \left(\frac{1}{\psi} - \frac{\psi+1}{\psi+S_i} - \frac{S_i}{(\psi+S_i)^2} \right) \right\} \right\} \psi + P_i \left(\frac{\psi}{\psi+S_i} \right)^{\psi} \left\{ \log \left(\frac{\psi}{\psi+S_i} \right) + \frac{S_i}{\psi+S_i} \right\} \right\} e^{\psi_1}
\end{aligned}$$

Figure 1.

Distribution of $\log(\text{PBB}+1)$ values in the MIPBB. The observed PBB levels range from less than 1.0 parts per billion (ppb) to 1114.2 ppb with a median of 2.0 ppb. More than 80% of the $\log(\text{PBB}+1)$ values are concentrated between $\log 1$ and $\log 11$. The shaded area represents the number of observations under the laboratory detection limit of 1.0 ppb.

Figure 2.

The solid line is the estimated density function of the log likelihood ratio from the simulated data with the setting $(I, n_i, \beta_0, \beta_1, \lambda, \psi, \gamma) = (1000, 4, 0.05, -0.01, 5.0, 0.1, 0)$ based on 1000 replicates. The other two lines correspond to the density functions for χ_0^2 and χ_1^2 .

Figure 3. Histogram of empirical Bayes estimates of $\log(w_i)$ superimposed with the density of $\log(w_i)$ for PVF and Gamma distributions. The solid line corresponds to $w_i \sim \text{PVF}(0.105, 0.162, 0.131)$ and the dotted line is for $w_i \sim \text{Gamma}(0.238, 0.238)$.

Table 1
Simulation results with 500 replicates and $(\beta_0, \beta_1, \lambda, \psi, \gamma) = (0.05, -0.01, 5.0, 0.1, 0.5)$ *

# of subjects	Parameter	True Value	Estimate	Est. SE	Emp. SD	Cov95
500	β_0	0.05	0.051	0.027	0.027	0.932
	β_1	-0.01	-0.01	0.005	0.005	0.954
	λ	5.0	4.999	0.118	0.123	0.934
	ψ	0.1	0.105	0.029	0.029	0.944
	γ	0.5	0.500	0.028	0.026	0.972
1000	β_0	0.05	0.051	0.019	0.018	0.960
	β_1	-0.01	-0.01	0.003	0.003	0.954
	λ	5.0	5.003	0.083	0.088	0.940
	ψ	0.1	0.103	0.020	0.020	0.964
	γ	0.5	0.499	0.020	0.019	0.968

*Model: $\log Y_{ij} = -\frac{1}{\lambda} \log w_i + \beta_0 + \beta_1 X_{ij} + \frac{1}{\lambda} e_{ij}$, where $w_i \sim \text{PVE}(\gamma, \psi)^{1-\gamma}$, $i = 1, \dots, K$; $j = 1, \dots, n_i$.

Table 2

The distribution of γ under the null hypothesis $\gamma = 0$ based on simulation with various sample sizes ($I=100, 500, 1000, 2000$). Here, $(n_i, \beta_0, \beta_1, \lambda, \psi, \gamma) = (4, 0.05, -0.01, 5.0, 0.1, 0)$ and results are based on 500 replicates.

Quant.	100	500	1000	2000
5%	0.000	0.000	0.000	0.000
95%	0.115	0.054	0.040	0.030
Mean	0.026	0.015	0.011	0.008
$Size^1$	0.032	0.026	0.022	0.020
$Size^2$	0.058	0.048	0.050	0.052

$Size^1$ Size under the χ^2_1 based likelihood ratio test.

$Size^2$ Size under the mixed chi-square based likelihood ratio test.

Table 3Results from the MIPBB [★]

(a) $w_i \sim \text{PVF}(\gamma, \psi^{1-\gamma}, \psi); -\log L = 4244.02$			
Variable	Parameter	Estimate (SE)	Z
Intercept	β_0	-0.338(0.027)	-12.52 *
Time, years	β_1	-0.002(0.001)	-2.00 *
(BMI-24), kg/m ²	β_2	-0.017(0.004)	-4.25 *
Time×(BMI-25)	β_3	0.001(0.0002)	5.00 *
	λ	5.991(0.168)	35.66 *
	ψ	0.131(0.017)	7.71 *
	γ	0.105(0.017)	6.18 *

(b) $w_i \sim \text{Gamma}(\psi, \psi); -\log L = 4259.72$			
Variable	Parameter	Estimate (SE)	Z
Intercept	β_0	-0.277(0.018)	-15.39 *
Time, years	β_1	-0.002(0.001)	-2.00 *
(BMI-24), kg/m ²	β_2	-0.016(0.004)	-4.00 *
Time×(BMI-25)	β_3	0.001(0.0002)	5.00 *
	λ	5.801(0.162)	35.81 *
	ψ	0.238(0.013)	18.31 *

[★]Model: $\log Y_{ij} = -\frac{1}{\lambda} \log w_i + \beta_0 + \beta_1 t_{ij} + \beta_2 (\text{BMI}_i - 25) + \beta_3 t_{ij} \times (\text{BMI}_i - 25) + \frac{1}{\lambda} e_{ij}$ where $i = 1, \dots, 955; j = 1, \dots, n_i$.

* P-value < 0.05 for the hypothesis test that the null value equals zero based on the Wald test.