Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Deng, et al. J Biomet Biostat 2011, 2:4
DOI: 10.4172/2155-6180.1000118

Research Article Open Access

Characterizing Components in a Mixture Model for Birthweight Distribution

Wei Deng[1], Richard Charnigo[2]*, Hongying Dai[3] and Russell S. Kirby[4]

[1]Departments of Biostatistics, University of Kentucky
[2]Departments of Statistics and Biostatistics, University of Kentucky
[3]Department of Medical Research, Children's Mercy Hospital
[4]Department of Community and Family Health, University of South Florida

Abstract

Low birthweight (LBW) is a well-known risk factor for infant mortality worldwide. Although infant mortality has decreased in the United States during the past 20 years, the incidence of LBW has increased, suggesting that further reductions in infant mortality may be possible if the incidence of LBW can be reduced. In the present work, we introduce a new analytic framework for revealing the relationships between latent variables representing components in a mixture model for birthweight distribution and various other risk factors. More specifically, we show how to estimate the probability that a risk factor is present within one of the mixture components as well as the probability of mixture component membership among infants for whom a risk factor is present, both at a fixed birthweight and averaged across birthweights. We illustrate our analytic framework using publicly available data for white singletons born in the United States between 1998 and 2002. This framework provides a quantitative approach for the prediction of how addressing a modifiable risk factor may affect both the incidence of LBW and infant mortality, thereby facilitating decision making regarding resource allocation toward addressing that risk factor.

Keywords: Birthweight Distribution; Infant Mortality; FLIC; Mixture Model; Parametric Mixtures of Logistic Regressions; Bayes' Theorem

Introduction

In both developed and developing countries, birthweight is arguably the single most important predictor of infant mortality, besides being significantly associated with infant and childhood morbidity [1-5]. Low birthweight (LBW, < 2500 g) infants have increased risk of developing cerebral palsy, hyaline membrane disease, apnoea, intracranial haemorrhage, sepsis, retrolental fibroplasia, and other conditions related to physiological immaturity [6-7].

Although infant mortality in the United States declined 45.2% from 1980 to 2000 (from 12.6 to 6.9 deaths per 1,000 live births), the percentage of LBW infants increased 11.8% (from 6.8 to 7.6 per 1,000 live births) and that of very low birthweight (VLBW, < 1500 g) infants increased 24.3% (from 1.15 to 1.43 per 1,000 live births) [8]. To the extent that LBW lies on a causal pathway leading to infant mortality, decreases in the former may further reduce the latter.

Several risk factors, both anthropic and environmental, have been implicated as possible contributors to LBW and, therefore, infant mortality. Based on a meta-analysis that included 17 observational studies, Vergnes and colleagues [9] identified 14 clusters of such risk factors while focusing on the role of periodontal disease on adverse pregnancy outcomes: 1. maternal age; 2. maternal general medical conditions; 3. maternal pregnancy associated conditions; 4. uterine, placental, or fetal abnormalities; 5. anthropometric factors; 6. socioeconomic status; 7. genitourinary tract infection; 8. other infection; 9. tobacco; 10. alcohol or drugs; 11. race or ethnicity; 12. prenatal care; 13. obstetric history; and, 14. dental treatment or oral hygiene.

The goal of the present work is to introduce a new analytic framework for revealing the relationships between such risk factors and the latent variables representing components in a mixture model for birthweight distribution. To make this article self-contained and to clarify this goal, we begin with a brief description of mixture modeling and its application to birthweight distribution.

Mixture Modeling and Birthweight Distribution

Mixture modeling is used to describe phenomena (such as birthweight) for which common parametric probability distributions are not suitable [10-14]; the citations in [14] identify some other relevant sources on mixture modeling. While a mixture model may be constructed from any family of common parametric probability distributions, in practice the normal family is most often used. A normal mixture model with k components is defined by the probability density function

$$\sum_{j=1}^{k} p_j f(x;\mu_j,\sigma_j), \tag{1}$$

where $f(x; \mu_j,\sigma_j)$, $1 \le j \le k$, is the probability density function of the normal distribution with mean and standard deviation σ_j, both positive numbers, while p_1 to p_k are nonnegative constants that sum to 1.

Notice that Equation (1) may include redundant components. For example, if $p_k = 0$ or if both $\mu_k = \mu_j$' and $\sigma_k = \sigma_j$', where j' is between 1 and k-1, then the k^{th} component is redundant. Moreover, the fact that such a redundancy may be achieved in two ways has profound

***Corresponding author:** Richard Charnigo, Departments of Statistics and Biostatistics, University of Kentucky, 851 Patterson Office Tower, University of Kentucky, Lexington, KY 40506-0027, Tel: 859.257.2550; Fax: 859.323.1973; E-mail: RJCharn2@aol.com

Received June 09, 2011; **Accepted** August 09, 2011; **Published** September 25, 2011

Citation: Deng W, Charnigo R, Dai H, Kirby RS (2011) Characterizing Components in a Mixture Model for Birthweight Distribution. J Biomet Biostat 2:118. doi:10.4172/2155-6180.1000118

implications for statistical inference. In particular, one cannot perform a chi-square-calibrated likelihood ratio test to determine the number of non-redundant components. To circumvent this difficulty, one may employ an information criterion to estimate the number of non-redundant components. Three such criteria are the FLIC [15], AIC [16], and BIC [17]. In the context of applying a normal mixture model to birthweight distribution, Charnigo and colleagues [15] argued in favor of using the FLIC; the FLIC is intermediate between the AIC and the BIC, and so the FLIC may avoid choosing either too many components (a potential weakness of the AIC) or too few (a potential weakness of the BIC). Alternatively, one may employ the approach of Li and Chen [18] for determining the number of non-redundant components. Their approach is based on sequential hypothesis testing underpinned by the expectation maximization algorithm. For the balance of this article, we will assume that all redundant components have been discarded, so that k in Equation (1) is the number of non-redundant components.

Equation (1) is often interpreted as saying that the full population consists of k subpopulations; the proportion of individuals in the full population belonging to subpopulation j is p_j, and in subpopulation j the measurements (birthweights, in this article) are normally distributed with mean μ_j and standard deviation σ_j. This interpretation comes with two key caveats. First, such subpopulations are not explicitly identified. In other words, subpopulation membership is a latent variable. Second, such subpopulations may not possess obvious biological meanings. Even so, when fitting two-component normal mixture models to birthweight distributions, Gage and Therriault [19] suggested that the two components might be roughly described as representing normal and abnormal developmental processes respectively.

Scope of the present work

The present work, however, goes beyond rough descriptions and explicitly relates mixture component membership to various risk factors for LBW and infant mortality, notwithstanding the two key caveats. Indeed, the present work is designed to mitigate these caveats. First, we will show how to obtain point and interval estimates of the probability that a risk factor is present within one of the mixture components as well as point and interval estimates of the probability of mixture component membership among infants for whom a risk factor is present, both at a fixed birthweight and averaged across birthweights. Thus, although mixture component memberships are not explicitly identified, probabilities can be assigned to the corresponding latent variable, thereby permitting informative statements such as, "A premature infant at birthweight 1600 g has a 99% chance of belonging to component 1." Second, while the mixture components themselves may lack obvious biological meanings, we will show how to probabilistically relate membership in the mixture components to factors that do have obvious biological meanings; for instance, the probability of membership in component 4 may be elevated when the infant's mother is diabetic.

The rest of this article is organized as follows. Section 2 presents our analytic framework, which combines a modified version of Gage's [20] parametric mixtures of logistic regressions [20,21] with Bayes' Theorem and the Law of Total Probability. Section 3 illustrates the analytic framework through application to data publicly available from the Centers for Disease Control and Prevention on white singletons born in the United States between 1998 and 2002. Section 4, besides identifying limitations of our analytic framework and opportunities for future research, describes how the analytic framework may be used to quantify the reductions in LBW incidence and infant mortality that may be attainable from interventions targeted toward a modifiable risk factor related to mixture component membership.

Analytic Framework

For ease of exposition, we present our analytic framework in seven steps.

Step 1

Acquire the data and classify the subjects. Draw with replacement N_{rep} samples of size n from the population of interest. For instance, the population of interest may be singleton infants born to non-Hispanic white mothers in the United States between 1998 and 2002 with known gestational ages of at least 22 weeks and known birthweights between 500 and 5500 g. Following Charnigo and colleagues [15], we suggest taking $N_{rep} = 25$ and $n = 50{,}000$. Of course, an investigator may choose N_{rep} and n however he/she wishes, subject to the following considerations: (i) taking N_{rep} too small may yield undesirably wide confidence intervals; (ii) there is a point of diminishing returns, depending on the amount of overlap among the samples, past which increasing N_{rep} may not substantially shorten confidence intervals; (iii) taking n too small may result in selection of a normal mixture model with too few components, diminishing the value of the subsequent analyses; and, (iv) taking n too large may immensely increase the computation time. Each infant in each sample is then classified as "high risk" or "low risk" on the factor being related to mixture component membership. For instance, if the factor is maternal diabetes, then infants born to diabetic mothers are classified as "high risk" and infants born to non-diabetic mothers are classified as "low risk".

Step 2

Select the number of components in the mixture model for birthweight distribution. This can be done by applying the FLIC to the birthweights in each of the N_{rep} samples [15]. The FLIC has the form $-2 \log L_k + 2(\log \sqrt{n})^{B(n,\delta)} (3k - 1)$, where L_k is the value of the likelihood function for the fitted k-component model and $B(n,\delta)$ is a function of the sample size and parameter estimates that falls between 0 and 1; one minimizes the FLIC over k between 1 and a prespecified upper bound such as 7 [15]. Then the number of components in the mixture model for birthweight distribution can be determined by a majority vote from the N_{rep} samples. For example, if the FLIC recommends a four-component mixture model for 23 out of $N_{rep} = 25$ samples, then a four-component mixture model can be adopted for the subsequent analyses.

Step 3

Given the selected number of components from Step 2, estimate the parameters in Equation (1) using each sample. Then combine the sample-specific estimates into overall estimates. The sample-specific estimates of parameters in Equation (1) may be obtained, in principle, by maximum likelihood. A practical implementation involving the expectation maximization algorithm and the optim function in the R statistical software package is described by Charnigo and colleagues [15]; the overall estimates are then acquired by averaging the sample-specific estimates. (These overall estimates are used in Steps 4 through 7 below whenever Equation (6) in [21] is invoked)

Step 4

Using each sample, estimate the probability of high risk classification on the factor being studied, as a function of both birthweight and component membership. Then combine the sample-

specific estimates into overall estimates and produce 95% confidence intervals. Symbolically, we need to estimate

P(high risk | birthweight, component), (2)

where P stands for probability, birthweight is continuous and ranges from (say) 500 to 5500 g, and component is categorical and ranges from 1 to k. Using any particular sample, this can be accomplished by applying a modified version of Gage's [20] parametric mixtures of logistic regressions (PMLR), as described by Charnigo and colleagues [21]. PMLR was originally designed as a tool to estimate birthweight-specific infant mortality within each component of a two-component normal mixture model but was subsequently modified to accommodate an arbitrary finite number of components. Importantly, modified PMLR is not the same as performing an ordinary logistic regression with birthweight and component membership allowed to interact; indeed, performing such a logistic regression is impossible anyway because component membership is a latent variable. Moreover, the present application of modified PMLR is distinguished from that in [21] in that now, instead of infant mortality, we take high risk classification on the factor being related to component membership as our dichotomous outcome variable. Once sample-specific estimates of Equation (2) are acquired, overall estimates and 95% confidence intervals may be obtained via formulas (6) and (7) in [21],

$\theta^* + \{ B^*_\theta + C S^*_\theta / \sqrt{N_{rep}} \}$ and $C = C_0 \sqrt{(\varphi N_{rep} / \{ 1 - (1- \varphi)^{Nrep} \})}$,

where θ^* is the mean of the sample-specific estimates, B^*_θ is a bias adjustment, S^*_θ is the standard deviation of the sample-specific estimates, C_0 is set to 4.0, and φ is set to n divided by the population size.

Step 5

Using each sample, estimate the probability of membership in each component, as a function of birthweight among those with high risk classification. Then combine the sample-specific estimates into overall estimates and produce 95% confidence intervals. Symbolically, we wish to estimate

P(component | birthweight, high risk). (3)

To do this, we employ Bayes' Theorem, which says that Equation (3) is equal to

P(high risk | birthweight, component) P(component | birthweight) / $\sum_{j=1}^{k}$ P(high risk | birthweight, j) P(j | birthweight).

Sample-specific estimates of quantities P(high risk | birthweight, j) have already been obtained in Step 4, and sample-specific estimates of quantities P(j | birthweight) may be acquired by substituting sample-specific estimates from Step 3 into the relation

P(j | birthweight) = p_j f(birthweight; μ_j,σ_j) / $\sum_{l=1}^{k}$ p_l f(birthweight; μ_l, σ_l).

Once sample-specific estimates of Equation (3) have been calculated, overall estimates and 95% confidence intervals may be obtained via formulas (6) and (7) in [21].

Step 6

Using each sample, estimate the probability of high risk classification, as a function of component membership. Then combine the sample-specific estimates into overall estimates and produce 95% confidence intervals. Symbolically, we need to estimate

P(high risk | component). (4)

To do so, we employ the Law of Total Probability, which says that Equation (4) is equal to

$\int$ P(high risk | birthweight, component) f(birthweight; $\mu_{component}$, $\sigma_{component}$) d(birthweight).

Sample-specific estimates of P(high risk | birthweight, component) are available from Step 4, and sample-specific estimates of f(birthweight; $\mu_{component}$, $\sigma_{component}$) can be derived from Step 3. Once sample-specific estimates of Equation (4) have been calculated, overall estimates and 95% confidence intervals may be obtained via formulas (6) and (7) in [21].

Step 7

Using each sample, estimate the probability of membership in each component, among those with high risk classification. Then combine the sample-specific estimates into overall estimates and produce 95% confidence intervals. Symbolically, we wish to estimate

P(component | high risk). (5)

To do this, we employ Bayes' Theorem, which says that Equation (5) is equal to

P(high risk | component) pcomponent /

$\sum_{j=1}^{k}$ P(high risk | j) p_j.

Sample-specific estimates of quantities P(high risk | j) have already been obtained in Step 6, and sample-specific estimates of quantities p_j have already been acquired in Step 3. Once sample-specific estimates of Equation (5) have been calculated, overall estimates and 95% confidence intervals may be obtained via formulas (6) and (7) in [21].

Empirical Investigation

To illustrate our analytic framework, we now describe an empirical investigation that we conducted using data publicly available from the Centers for Disease Control and Prevention via {http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm}. More specifically, we used United States birth-cohort-linked infant birth and death data (and fetal death data) from years 1998 to 2002, subject to the following inclusion/exclusion criteria. First, we limited the maternal race to white, since relationships between mixture component membership and risk factors may differ by race; in fact, because there are documented racial differences in birthweight distribution [8], even the appropriate number of mixture components may depend on race. Likewise, we restricted attention to non-Hispanic mothers. Second, we included only records with known gestational ages of at least 22 weeks and known birthweights between 500 and 5500 g, as gestational ages less than 22 weeks or birthweights outside the indicated range may not be accurately documented [22]; we also omitted records with missing values on one or more of 16 variables in an initial list of risk factors. Third, we restricted attention to singleton births. This not only avoids a source of heterogeneity, inasmuch as multiple births tend to have smaller birthweights and gestational ages than singleton births [23,24], but also circumvents the problem that observations on multiple births are not statistically independent; our analytic framework assumes that observations within each sample are statistically independent.

The above inclusion/exclusion criteria yielded an effective population size of 4,034,402, from which we drew N_{rep} = 25 samples

of size n = 50,000. The FLIC recommended a four-component mixture model for 23 of the samples, so that a four-component mixture model was adopted for the subsequent analyses. The overall estimates for the parameters in this model defined the probability density function

$$0.004\, f(x; 908{,}249) + 0.112\, f(x; 2932{,}734) + 0.820\, f(x; 3416{,}427) + 0.065\, f(x; 4044{,}431). \quad (6)$$

Thus, for example, if infants with a particular risk factor for LBW and infant mortality have an estimated probability of belonging to component 2 that is substantially greater than 0.112, then we may say that such infants have excess membership in component 2 (compared to infants more generally in the population of interest).

Following exploratory analyses of the first two samples in which point estimates of Equations (4) and (5), but not confidence intervals, were obtained for each variable in our initial list of 16 risk factors (results available from the corresponding author upon request), we narrowed our initial list of 16 risk factors to four risk factors that we then related to membership in the mixture model's four components. The four risk factors were: (i) gestational age with a cutpoint of 29 weeks, strictly below which infants were severely premature and classified as high risk on that factor; (ii) maternal diabetes (including juvenile onset, adult onset, and gestational), the presence of which defined high risk on that factor; (iii) maternal chronic high blood pressure (HBP) (including diagnosis prior to pregnancy or before the 20th week of gestation), the presence of which defined high risk; and, (iv) maternal previous preterm birth, the presence of which defined high risk.

(Figures 1-4) Depict the estimated probabilities of high risk classifications on the four risk factors, as functions of both birthweight and component membership, along with (pointwise) 95% confidence bounds. Scaling on the vertical axes is logarithmic, and the curves are suppressed at birthweights more than three component standard deviations away from the corresponding component mean. In general, the confidence bounds tend to be wider at extreme birthweights than at relatively normal birthweights; this occurs because there are many

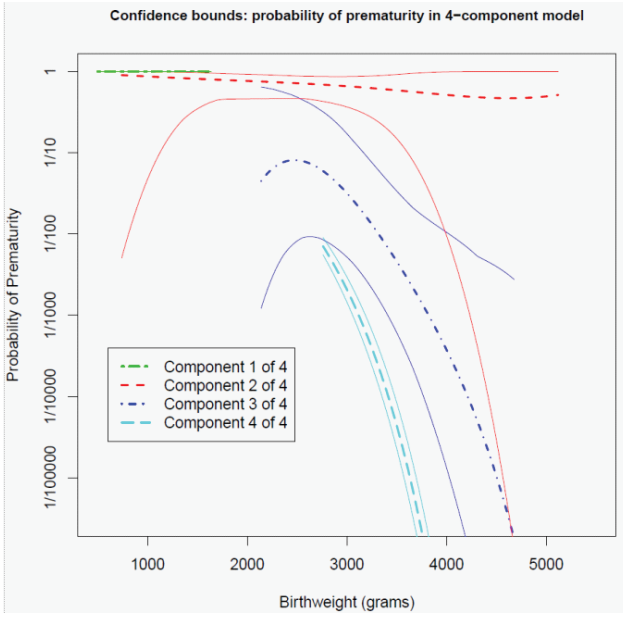

Figure 1: Estimated probabilities of severe prematurity within each component as a function of birthweight. Displayed are estimated probabilities of severe prematurity (< 29 weeks) within each component as a function of birthweight, along with 95% confidence intervals. (Table 1) shows the estimated probabilities of severe prematurity within each component, averaged over birthweight in the sense of Equation (4).

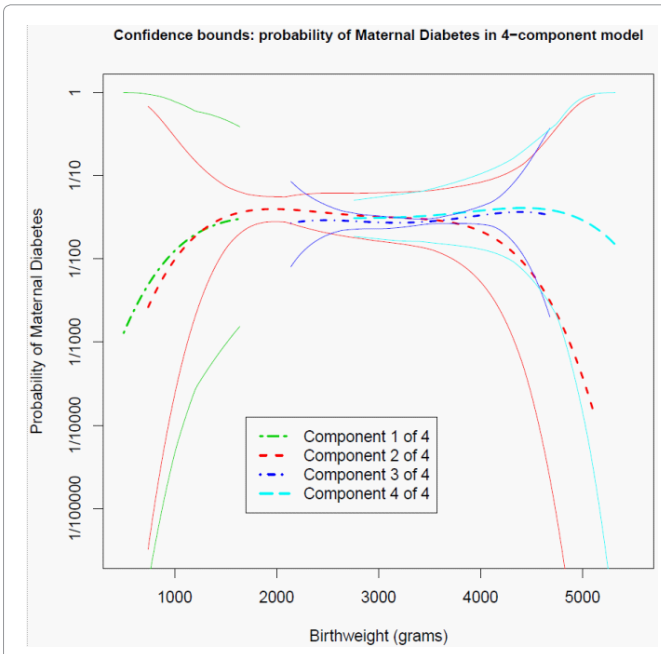

Figure 2: Estimated probabilities of maternal diabetes within each component as a function of birthweight. Displayed are estimated probabilities of maternal diabetes within each component as a function of birthweight, along with 95% confidence intervals. (Table 1) shows the estimated probabilities of maternal diabetes within each component, averaged over birthweight in the sense of Equation (4).

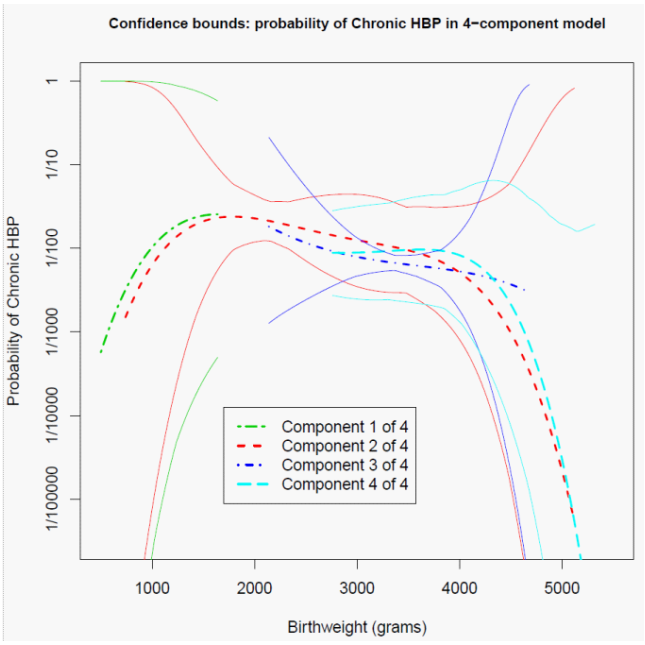

Figure 3: Estimated probabilities of maternal chronic HBP within each component as a function of birthweight. Displayed are estimated probabilities of maternal chronic HBP within each component as a function of birthweight, along with 95% confidence intervals. (Table 1) shows the estimated probabilities of maternal chronic HBP within each component, averaged over birthweight in the sense of Equation (4).

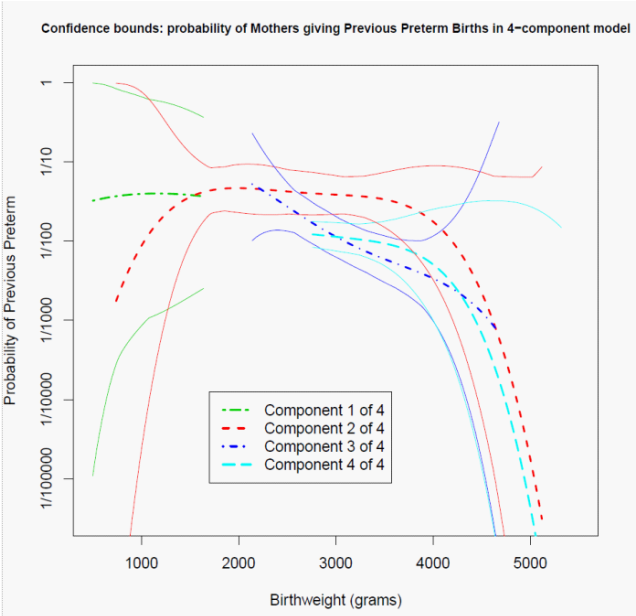

Figure 4: Estimated probabilities of a previous preterm birth within each component as a function of birthweight.
Displayed are estimated probabilities of a previous preterm birth within each component as a function of birthweight, along with 95% confidence intervals. (Table 1) shows the estimated probabilities of a previous preterm birth within each component, averaged over birthweight in the sense of Equation (4).

Component	Est. Probability of Prematurity (per 1,000 births)	95% Confidence Interval	
		Lower	Upper
1	1000.0	1000.0	1000.0
2	636.7	325.0	864.5
3	15.7	2.9	79.3
4	0.1	0.0	1.5
Component	Est. Probability of Maternal Diabetes (per 1,000 births)	95% Confidence Interval	
		Lower	Upper
1	14.8	0.4	354.0
2	34.2	20.0	57.8
3	29.5	25.8	33.7
4	42.3	16.0	106.9
Component	Est. Probability of Maternal Chronic HBP (per 1,000 births)	95% Confidence Interval	
		Lower	Upper
1	16.4	0.1	777.7
2	15.6	8.4	29.0
3	7.1	5.4	9.4
4	6.5	1.1	37.6
Component	Est. Probability of Previous Preterm Birth (per 1,000 births)	95% Confidence Interval	
		Lower	Upper
1	46.9	5.2	315.1
2	38.3	23.0	63.0
3	8.5	5.1	14.0
4	5.5	1.8	16.7

Table 1: Estimated probabilities of risk factors within each component.
Displayed are estimated probabilities of severe prematurity (< 29 weeks), maternal diabetes, maternal chronic HBP, and a previous preterm birth within each component, along with 95% confidence intervals.

more infants with relatively normal birthweights, so that probabilities of high risk classifications can be estimated more precisely for such infants.

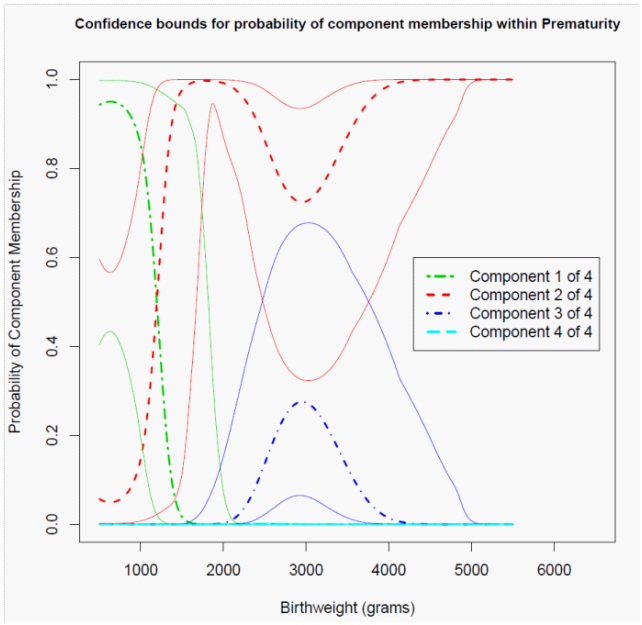

Figure 5: Estimated probabilities of component membership as a function of birthweight among severely premature infants. Displayed are estimated probabilities of component membership as a function of birthweight among severely premature infants (< 29 weeks), along with 95% confidence intervals. (Table 2) shows the estimated probabilities of component membership among severely premature infants, averaged over birthweight in the sense of Equation (5).

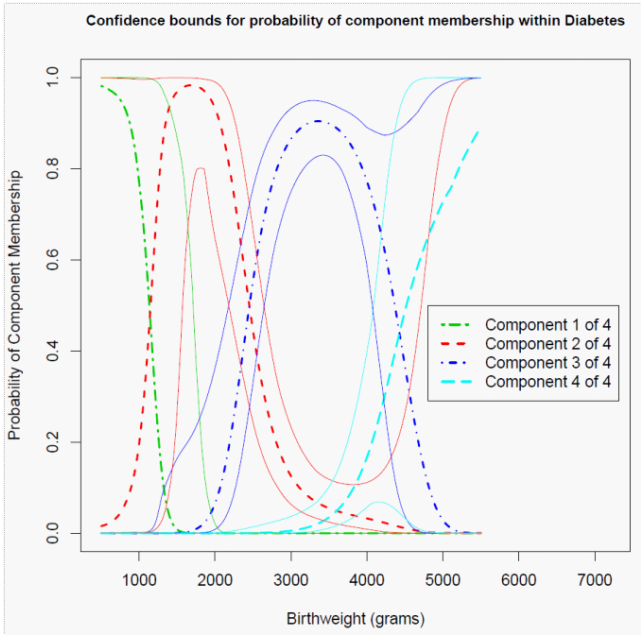

Figure 6: Estimated probabilities of component membership as a function of birthweight among infants with diabetic mothers. Displayed are estimated probabilities of component membership as a function of birthweight among infants with diabetic mothers, along with 95% confidence intervals. (Table 2) shows the estimated probabilities of component membership among infants with diabetic mothers, averaged over birthweight in the sense of Equation (5).

(Table 1) Presents the estimated probabilities of high risk classifications, as functions of component membership and averaged over birthweight per Equation (4), along with 95% confidence intervals. Of particular interest are the overlaps among the 95% confidence intervals for a given risk factor; the less overlap there is, the more indication we have that the risk factor is associated with component membership.

Gestational age is very strongly related to mixture component membership, as severe prematurity is ubiquitous within component 1, common within component 2, rare within component 3, and almost nonexistent within component 4. Maternal diabetes and maternal chronic HBP are less obviously related to mixture component membership. There is some suggestion that maternal diabetes may be most common in component 4 and least common in component 1, with an opposite pattern for maternal chronic HBP. However, definitive conclusions are impeded by wide confidence intervals for components 1 and 4, which are in turn related to the relatively small proportions of infants in those components. The presence of a previous preterm birth is clearly related to mixture component membership, as indicated by separation of the component 2 confidence interval from the component 3 and component 4 confidence intervals. In addition, although the component 1 confidence interval is too wide to permit a definitive conclusion, there is some suggestion that the presence of a previous preterm birth may be even more prevalent in component 1 than in component 2.

(Figures 5-8) Depict estimated probabilities of component membership, as functions of birthweight among those with high risk classifications, along with 95% confidence bounds. (Table 2) Presents estimated probabilities of component membership, among those with

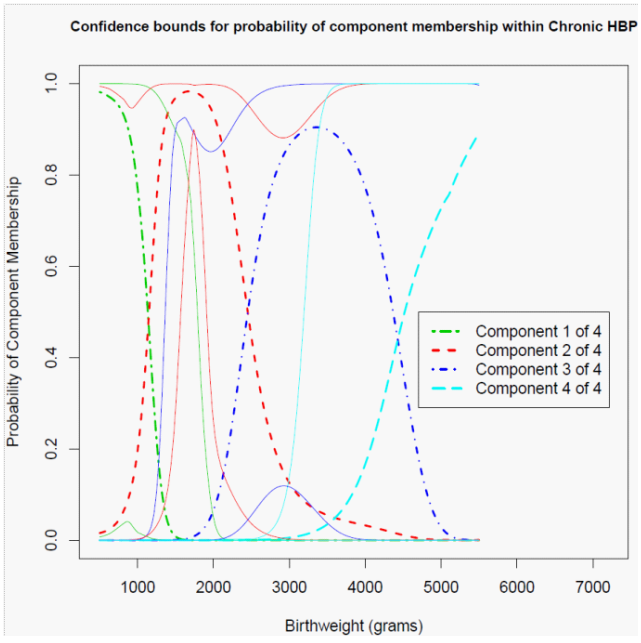

Figure 7: Estimated probabilities of component membership as a function of birthweight among infants with mothers who have chronic HBP.Displayed are estimated probabilities of component membership as a function of birthweight among infants with mothers who have chronic HBP, along with 95% confidence intervals. Table 2 shows the estimated probabilities of component membership among infants with mothers who have chronic HBP, averaged over birthweight in the sense of Equation (5).

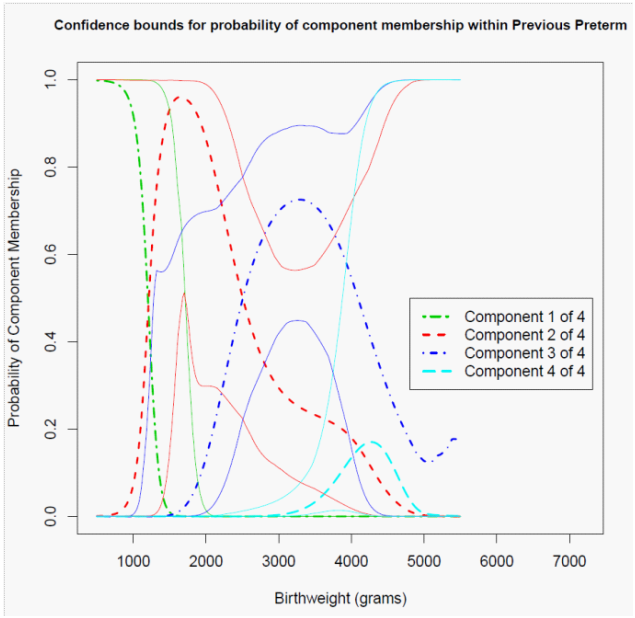

Figure 8: Estimated probabilities of component membership as a function of birthweight among infants with mothers who have a previous preterm birth. Displayed are estimated probabilities of component membership as a function of birthweight among infants with mothers who have a previous preterm birth, along with 95% confidence intervals. (Table 2) shows the estimated probabilities of component membership among infants with mothers who have a previous preterm birth, averaged over birthweight in the sense of Equation (5).

Component	Est. Probability of Membership among Premature (per 1,000 Births)	95% Confidence Interval	
		Lower	Upper
1	45.7	25.0	82.0
2	782.0	538.0	917.0
3	149.0	31.2	487.8
4	0.0	0.0	1.6
Component	Est. Probability of Membership among Maternal Diabetic (per 1,000 Births)	95% Confidence Interval	
		Lower	Upper
1	2.1	0.1	52.6
2	120.3	69.4	200.4
3	772.3	669.0	850.6
4	85.1	29.3	222.9
Component	Est. Probability of Membership among Maternal Chronic HBP (per 1,000 Births)	95% Confidence Interval	
		Lower	Upper
1	7.8	0.0	612.7
2	201.3	94.0	379.7
3	690.9	503.0	831.6
4	59.3	9.3	298.0
Component	Est. Probability of Membership among Previous Preterm Birth (per 1,000 Births)	95% Confidence Interval	
		Lower	Upper
1	15.1	2.0	105.5
2	350.5	175.5	577.6
3	579.5	359.7	771.8
4	26.8	5.7	117.9

Table 2: Estimated probabilities of component membership among those with risk factors.
Displayed are estimated probabilities of membership within each component among infants born severely premature (< 29 weeks), to mothers with diabetes, to mothers with chronic HBP, and to mothers with a previous preterm birth, along with 95% confidence intervals.

high risk classifications and averaged over birthweight per Equation (5), along with 95% confidence intervals.

Even though severe prematurity is more common within component 1 than within component 2, the overall proportion of infants in component 2 is sufficiently larger than the overall proportion of infants in component 1 that a severely premature infant is most likely to belong to component 2. Membership in component 3 is not particularly common for a severely premature infant but is not especially rare either, inasmuch as the overall proportion of infants in component 3 is larger than the overall proportion of infants in all other components combined. However, a severely premature infant will almost never belong to component 4. Infants born to diabetic mothers and infants born to mothers with chronic HBP are most likely to belong to component 3, although the former infants appear to have a modest excess in component 4 while the latter infants seem to have a slight excess in component 2. Infants born to mothers with a previous preterm birth are most likely to be in component 2 or component 3, with a modest excess in component 1 and a slight dearth in component 4.

Discussion

This article presented a new analytic framework for exhibiting relationships between latent variables representing components in a mixture model for birthweight distribution and various risk factors for LBW and infant mortality. Our framework builds upon the methodology of Charnigo and collaborators, who addressed the questions of how to choose the number of components in the mixture model [15] and how to estimate infant mortality within each component [21]. However, our efforts go beyond theirs in that we explicitly characterize the mixture components in terms of observable risk factors with biological meaning. In particular, we indicate how to estimate the probability that a risk factor is present within one of the mixture components as well as the probability of mixture component membership among infants for whom a risk factor is present.

While our characterizations of the mixture components are probabilistic rather than deterministic, they nonetheless provide insight into how decision makers might best allocate resources for educational activities and other interventions directed at women who are or who may become pregnant. Indeed, the analyses presented in this article for white singletons regarding gestational age, maternal diabetes, maternal chronic HBP, and previous preterm birth can be repeated for other populations and regarding other risk factors, including risk factors that are modifiable (such as maternal tobacco use, alcohol use, infrequent prenatal visits, and poor oral hygiene, to name a few). If the presence of one modifiable risk factor substantially increases the probability of membership in a component with a higher incidence of LBW or infant mortality, while the presence of another modifiable risk factor does not, then - all else being equal - the former modifiable risk factor is a more suitable target for interventions.

For example, suppose that infant mortality is 100/20/6/8 per 1,000 births in the four components respectively and that component membership probabilities are 1/15/75/9 when modifiable risk factor "A" is present, 0/7/90/3 when "A" is absent, 1/14/78/7 when modifiable risk factor "B" is present, and 0/8/87/5 when "B" is absent. Then infant mortality is estimated to be 9.22 per 1,000 births when "A" is present, 7.04 when "A" is absent, 9.04 when "B" is present, and 7.22 when "B" is absent. So, if interventions targeting "A" and "B" are equally costly and equally effective at removing the modifiable risk factor, then the intervention for "A" is favored because of its greater impact on infant mortality. In fact, the preceding computations are conservative since removing "A" or "B" may impact infant mortality in two ways: by shifting component membership probabilities toward favorable components (in the sense of lower infant mortality), which is incorporated into the above calculations, and by reducing infant mortality within any given component, which is not built into the above computations and which will be a topic for future research. Of course, one may also perform similar calculations for LBW, preterm birth, or small-for-gestational age, rather than for infant mortality.

Our empirical investigation has several limitations. Some limitations were inherited from the data and were not issues with our analytic framework per se. The samples we drew might have reflected a selection bias since missing values might not have occurred at random on some risk factors. Moreover, some variables potentially related to mixture component membership, such as maternal oral hygiene, were not available from the data. Additionally, imperfections in documenting diabetes and hypertension on birth certificates may have obscured these risk factors' associations with component membership; a way to overcome this limitation in future research may be to use a linked birth certificate-maternal hospital discharge database [25].

One limitation was more conceptual in nature. The FLIC led us to a four component mixture model, and we then related those four components to observable risk factors with biological meaning. However, a researcher does well to remember the distinction between a statistical model and the natural phenomenon being described by it; in particular, the four component mixture model is only a mathematical approximation to a birthweight distribution. Even so, the four component mixture model may prove useful in reducing the incidence of LBW and infant mortality, by informing the allocation of resources for educational activities and other interventions targeting modifiable risk factors. If there had been ambiguity about the appropriate number of components, in that the FLIC had recommended (say) a four-component model for 13 samples and a six-component model for 12 samples, then we could have performed the subsequent analyses for both a four-component model and a six-component model. However, if applied to estimate the potential benefits of educational activities and other interventions as in the hypothetical example with modifiable risk factors "A" and "B", four-component and six-component models might yield different conclusions. In that case, one might calculate a weighted average of the estimated potential benefits based on the two models, the weight for each model proportional to the number of times that model had been recommended by the FLIC.

Other limitations were analytical; in fact, these limitations suggest the next steps for methodological development. First, we treated the risk factors as dichotomous. However, many risk factors are naturally continuous (or, at least, interval level); a statistical model that could accommodate the gradations inherent to such risk factors might be more realistic and might therefore better serve the ultimate goal of reducing the incidence of LBW and infant mortality. Second, we examined each risk factor separately. A desirable pursuit for future research would be to assess multiple risk factors simultaneously, so that we could estimate, for example, the probabilities of mixture component membership for infants whose mothers had both diabetes and chronic HBP but not a previous preterm birth. Third, a valuable task for future research would be to explicitly incorporate infant mortality, birthweight, and the risk factor(s) into a single, unified statistical model. This would permit estimating the full impact on infant mortality of removing a modifiable risk factor, not just the contributions from shifting component membership probabilities toward favorable components but also the benefits from reducing infant mortality within any given component.

Finally, while the analytic framework described in this article was motivated by the task of characterizing components in a mixture model for birthweight distribution (and the ultimate goal of reducing the incidence of LBW and infant mortality), one may apply the statistical tools from this analytic framework to problems in other biological arenas that employ mixture modeling, such as describing the differential expression of genes among patients afflicted with an illness versus healthy controls [26].

References

1. McCormick MC (1985) The Contribution of Low Birth weight to Infant Mortality and Childhood Morbidity. N Engl J Med 312: 82-90.
2. Yerushalmy J, Vandenberg BJ, Erhardt CL, Jacobziner H (1965) Birth weight and Gestation as Indices of "immaturity": Neonatal Mortality and Congenital Anomalies of the "immature". Am J Dis Child 109: 43-57.
3. Lubchenco L, Searls DT, Brazie JV (1972) Neonatal Mortality Rate: Relationship to Birth weight and Gestational Age. J Pediatr 81: 814-822.
4. Hoffman HJ (1977) Classification of Births by Weight and Gestational Age for Future Studies of Prematurity. Epidemiology of Prematurity 297-333.
5. Semenciw RM, Morrison HI, Lindsay J, Silins J, Sherman GJ, et al. (1986) Risk Factors for Postneonatal Mortality: Results from a Research Linkage Study. Int J Epidemiol 15: 369-372
6. Ellenberg JH, Nelson KB (1979) Birth weight and Gestational Age in Children with Cerebral Palsy or Seizure Disorders. Am J Dis Child 133: 1044-1048.
7. Kramer MS (1987) Determinants of Low Birthweight: Methodological Assessment and Meta-analysis. Bull World Health Organ 65: 663-737
8. Iyasu S, Tomashek K, Barfield W (2002) Infant Mortality and Low Birth weight Among Black and White Infants --United States, 1980-2000. MMWR Morb Mortal Wkly Rep 51: 589-592
9. Vergnes JN, Sixou M (2007) Preterm Low Birth weight and Maternal Periodontal Status: A Meta-analysis. Am J Obstet Gynecol 196: 135 e1-7.
10. Lindsay BG (1995) Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics 5.
11. Chen H, Chen J, Kalbfleisch JA (2001) Modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society Series B. 63: 19-29.
12. Charnigo R, Pilla RS (2007) Semiparametric Mixtures of Generalized Exponential Families. Scandinavian Journal of Statistics. 34: 535-551.
13. Dai H, Charnigo R (2008) Inferences in Contaminated Regression and Density Models. Sankhya. 69: 842-869.
14. Charnigo R, Sun J (2010) Asymptotic Relationships between the D-Test and Likelihood Ratio-Type Tests for Homogeneity. Statistica Sinica 20: 497-512.
15. Charnigo R, Chesnut LW, LoBianco T, Kirby RS (2010a) Thinking Outside the Curve, Part I: Modeling Birthweight Distribution. BMC Pregnancy and Childbirth 10: 10:37.
16. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F eds Akademiai Kiado, Budapest Second International Symposium on Information Theory.
17. Schwarz G Estimating the dimension of a model. Annals of Statistics 6: 461-464
18. Li P, Chen J (2010) Testing the order of a finite mixture model. Journal of the American Statistical Association 105: 1084-1092.
19. Gage TB, Therriault G (1998) Variability of Birth-Weight Distributions by Sex and Ethnicity: Analysis Using Mixture Models. Human Biology 70: 517- 534.
20. Gage TB (2002) Birth-Weight-Specific Infant and Neonatal Mortality: Effects of Heterogeneity in the Birth Cohort. Human Biology 81: 753-772.
21. Charnigo R, Chesnut LW, LoBianco T, Kirby RS (2010b) Thinking Outside the Curve, Part II: Modeling Fetal-infant Mortality. BMC Pregnancy and Childbirth 10:44
22. MacDorman MF, Mathews TJ (2009) Behind International Rankings of Infant Mortality: How the United States Compares with Europe. NCHS Data Brief 23: 1-8
23. Luke B (1996) Reducing fetal deaths in multiple births: optimal birthweights and gestational ages for infants of twin and triplet births. Acta Genet Med Gemellol (Roma) 45: 333-348.
24. Alexander G, Wingate MS, Salihu H, Kirby RS (2005) Fetal and Neonatal Mortality Risks of Multiple Births. Obstetrics and Gynecology Clinics of North America 32: 1-16.
25. Lydon-Rochelle MT, Holt VL, Cardenas V, Nelson JC, Easterling TR, et al. (2005) The reporting of pre-existing maternal medical conditions and complications of pregnancy on birth certificates and in hospital discharge data. American Journal of Obstetrics and Gynecology 193: 125-134.
26. Dai H, Charnigo R (2010) Contaminated Normal Modeling with Application to Microarray Data Analysis. Canadian Journal of Statistics. 38: 315-332.