Published in final edited form as:

Epidemiology. 2018 March ; 29(2): 183–190. doi:10.1097/EDE.0000000000000789.

Hierarchical semi-Bayes methods for misclassification in perinatal epidemiology

Richard F MacLehose¹, Lisa M Bodnar², Craig Meyer¹, Haitao Chu³, and Timothy L Lash⁴

¹Division of Epidemiology and Community Health, University of Minnesota

²Departments of Epidemiology and of Obstetrics, Gynecology, and Reproductive Sciences, University of Pittsburgh

³Division of Biostatistics, University of Minnesota

⁴Department of Epidemiology, Rollins School of Public Health, Emory University

Abstract

Background—Validation data are used to estimate the extent of misclassification in epidemiologic studies. In the Penn MOMS cohort, prepregnancy BMI is subject to misclassification and validation data are available to estimate the extent of misclassification. We use these data to estimate the association between maternal prepregnancy body mass index and early preterm (<32 weeks) birth using a semi-Bayes hierarchical model allowing for more flexible adjustment for misclassification.

Methods—We propose a two-stage model that first fits a Bayesian hierarchical model for the bias parameters in the validation study. This model shrinks bias parameters in different groups toward one another in an effort to gain precision and improve mean squared error. In the second stage, we draw random samples from the posterior distribution of the bias parameters to implement a probabilistic bias analysis adjusting for exposure misclassification in a frequentist outcome model.

Results—Bias parameters from the hierarchical model were often more substantively reasonable and often had smaller variance. Adjusting results for misclassification generally attenuated the strength of the unadjusted associations. After adjusting for misclassification, underweight mothers were not at increased risk of early preterm birth relative to normal weight mothers. Severely obese mothers had an increased risk of early preterm birth relative to normal weight mothers.

Conclusions—The two-stage semi-Bayesian hierarchical model borrowed strength between group-specific bias parameters to adjust for exposure misclassification. Model results support evidence of an increased risk of early preterm birth among severely obese mothers, relative to normal weight mothers.

Corresponding author: Richard F. MacLehose, Ph.D, Division of Epidemiology, 1300 S 2nd St #300, Minneapolis, MN 55454, Phone: (612) 624-1932, mac10029@umn.edu.

Conflicts of Interest: None

Code is included as an e-appendix.

Keywords

Bias Analysis; Bayes; Hierarchical; Uncertainty Analysis; Maternal BMI; preterm birth

Introduction

Observational research is often susceptible to misclassification of exposure, disease, and confounders. The previous decade has seen an abundance of research on methods to adjust estimates of association for misclassification and other types of bias.^{1–4} Most of the methods require specifying a bias model that includes parameters such as the positive and negative predictive values of classification. These parameters influence the magnitude and direction of the bias adjustment, yet are not identified in the observed data. While prior studies can help inform assignment of credible values to the bias model parameters,^{5–8} some studies include validation substudies to estimate these bias parameters in the study population of interest.⁹

One approach to adjust for misclassification, when validation data are available, is to estimate the bias parameters (e.g., predictive values) and their variances from the validation data and use them to impute the data that would have been observed in the absence of misclassification.⁹ We note that while classification probabilities (sensitivity and specificity) are typically used in bias analysis, predictive probabilities are more readily measured in validation studies.¹⁰ When an exposure or covariate is the misclassified variable of concern, these values are often estimated separately for those with and without the outcome. However, they are assumed to be constant across strata of all other covariate patterns. Although it may be substantively reasonable to suspect that the predictive values could depend on other factors, estimating stratum specific predictive values may be impractical due to sparse data.

Hierarchical models have a long history as a method to stabilize parameter estimates in the presence of sparse data.^{11,12} These models can estimate separate parameter values for each specified stratum while simultaneously borrowing information across other strata.¹³ In simple hierarchical models, stratum-specific estimates are a weighted average of the data in that stratum and the overall grand-mean estimate across all strata.¹⁴ Parameter values estimated from hierarchical models can have lower mean squared error than parameter values estimated by conventional methods that separately analyze each stratum of the data.¹⁵

Hierarchical models may be advantageous in estimating predictive values from validation data—and for modeling bias parameters more generally—and have been implemented previously by Greenland.² Validation studies may not have enough data to reliably estimate predictive values separately within all strata of interest with suitable precision. By borrowing information from other strata, however, hierarchical models may make estimation of these predictive values more tenable. Although hierarchical models are relatively straightforward and have notable advantages, their implementation in the context of bias analysis may increase computational time and require additional technical expertise. A fully Bayesian approach to hierarchical modeling of validation data in the context of exposure misclassification would require simultaneously fitting 1) an outcome model that estimates

the effect of the misclassification-adjusted exposure on the outcome of interest and 2) a bias parameter model that fits a hierarchical model of the predictive probability of true exposure. Further, each model would include prior distributions on all random variables.

Unfortunately, the complexity of this joint model specification can lead to difficulties in model convergence and can be very computationally time-intensive. Fortunately, there is generally little information contained in the outcome model that can aid in the estimation of the bias parameter model.^{2,16} We therefore follow Greenland² in proposing a two-step semi-Bayesian approach that decouples the two models.

The purpose of this study is to propose and implement a semi-Bayes hierarchical model to allow for more flexible implementation of adjustments for misclassification. We first fit a Bayesian hierarchical model of the bias parameters, and then use random samples from the posterior distributions of those parameters to implement a probabilistic bias analysis for the frequentist outcome model. We implement this approach in the Penn MOMS cohort to estimate the association between maternal prepregnancy body mass index (BMI) and early preterm birth.¹⁷ In this study, prepregnancy weight and height on birth certificate are recalled after delivery and are subject to misclassification, but validation data based on medical record review are available to inform predictive values.¹⁸

Methods

Penn MOMS Cohort and Validation Substudy

Penn MOMS is a birth certificate-based study of maternal prepregnancy BMI and gestational weight gain in relation to poor birth outcomes.¹⁷ The University of Pittsburgh IRB deemed the study not human subjects research due to lack of identifiable data. The cohort included all singleton infants born to non-Hispanic White or non-Hispanic Black women in Pennsylvania from 2003 to 2010 with complete data on key variables (n=773,625).

The birth certificate-derived prepregnancy BMI was defined as self-reported prepregnancy weight (converted from pounds to kilograms) divided by self-reported height (converted from inches to meters) squared, and was available for all participants. Gestational age at delivery was based on the best obstetric estimate as provided on the birth certificate. We defined early preterm birth as delivery at <32 weeks of gestation.

We validated prepregnancy BMI by sampling records from the eligible Penn MOMS cohort who were born at Magee-Womens Hospital in Pittsburgh. We hypothesized that the accuracy of prepregnancy BMI categorization on the birth certificate would vary according to gestational weight gain, maternal race/ethnicity, and gestational age at delivery. To permit estimation in rare groups, we used a balanced design.¹⁹ Up to 30 birth certificates were randomly sampled from those available in 48 strata created by simultaneous stratification on four birth certificate variables: prepregnancy BMI (underweight [<18.5 kg/m²], normal weight/overweight [18.5 – 29.9 kg/m²], obese class 1 [30 – 34.9 kg/m²], severely obese [>35 kg/m²]), gestational weight gain (<20th percentile, 20–80th percentile, >80th percentile), race/ethnicity (non-Hispanic white, non-Hispanic black), and gestational age (term, preterm

(<37 weeks)), as described previously.^{18,20} Medical charts of the sampled births were reviewed by trained abstractors to ascertain maternal report at the initial prenatal visit of prepregnancy weight and height. The gold-standard prepregnancy BMI was calculated and categorized using the medical record information.

Statistical Methods

The statistical analysis combines two models: 1) a Bayesian hierarchical model of the probability of an individual's true prepregnancy BMI being in a given category, conditional on the reported prepregnancy BMI category and the covariate stratum of interest, and 2) a frequentist model of the probability of early preterm birth conditional on the true prepregnancy BMI category and potential confounders. The Bayesian hierarchical model for the true prepregnancy BMI category is based on data collected in the validation substudy. The true prepregnancy BMI status, $T_{r,g,s,p}$ is a categorical variable that can take values 1 (underweight), 2 (normal weight), 3 (overweight), 4 (obese), or 5 (severely obese) and, for women in the validation substudy, is known from medical record review. We have opted to treat BMI as a categorical, rather than continuous variable, because this is of greatest substantive interest and is a common approach with this type of data. The subscript r indexes self-reported race (non-Hispanic White or non-Hispanic Black); g indexes the gestational weight gain category (<20th, 20th–80th, >80th); s indexes the birth certificate-reported prepregnancy BMI category based on birth certificate data (taking values from 1–5), p indexes preterm or term delivery. We aggregate women within categories and assume the aggregate random variable $T_{r,g,s,p}$ follows a multinomial distribution:

$$T_{r,g,s,p} \sim \text{Multinomial}(p_{r,g,s,p}, N_{r,g,s,p})$$

where $N_{r,g,s,p}$ is the total number of participants in the validation substudy of a given race/ethnicity, gestational weight gain category, birth certificate reported prepregnancy BMI category, and term/preterm status. The parameter

$p_{r,g,s,p} = [p_{r,g,s,p}^1, p_{r,g,s,p}^2, p_{r,g,s,p}^3, p_{r,g,s,p}^4, p_{r,g,s,p}^5]$ is a vector containing the probability a woman is truly in prepregnancy BMI category 1, 2, 3, 4 or 5 (the superscripts) given her birth certificate reported prepregnancy BMI category, gestational weight gain category, race/ethnicity, and term/preterm status.

Various approaches exist for modeling this type of data. One approach is to estimate the predictive probabilities separately for each stratum of the covariates, in which case the maximum likelihood estimate is $\hat{p}_{r,g,s,p}^j = t_{r,g,s,p}^j / N_{r,g,s,p}$, where the numerator of this quantity is the number of people with the j th prepregnancy BMI level conditional on reported prepregnancy BMI category, gestational weight gain category, race/ethnicity and term/preterm status. In this validation substudy, $N_{r,g,s,p}$ had an observed range of 2 to 34, and 14 of the 60 strata had $N_{r,g,s,p} = 10$. The maximum likelihood estimates of predictive probabilities in these sparse-data strata would be very imprecise due to the small sample sizes used to estimate them. An alternative approach would aggregate data over categories of gestational weight gain, and race/ethnicity. The aggregated data could be used to estimate one common set of classification probabilities for all women who self-report a given

prepregnancy BMI category ignoring the covariate strata: $\hat{p}_{s,p}^j = \sum_{r,g} t_{r,g,s,p}^j / \sum_{r,g} N_{r,g,s,p}$. This pooled estimate would have improved precision, relative to the stratum-specific estimates, but requires the questionable assumption that the predictive values are identical across women of different gestational weight gain, and race/ethnicity categories.

As a compromise between aggregating over all groups and keeping all groups completely distinct, we implemented a hierarchical regression model that estimates predictive values for each stratum while also borrowing information from other strata to stabilize estimates. The multinomial nature of the data requires special care when specifying the hierarchical structure. One approach to modeling this type of data structure is to place a prior distribution directly on the vector of classification probabilities, $p_{r,g,s,p}$. A Dirichlet distribution is a common choice in this situation; however, we found that this prior distribution led to model convergence problems. We opted instead to follow a hierarchical multinomial modeling approach similar to that proposed by Gelman et al.²¹ Rather than directly model the predictive probabilities, we first expressed the probabilities in terms of redundant variables, $\phi_{r,g,s,p}^i$ where

$$p_{r,g,s,p}^i = e^{\phi_{r,g,s,p}^i} / \sum_i e^{\phi_{r,g,s,p}^i} \quad (1)$$

The hierarchical specification will be based on the redundant parameters because, unlike the predictive values, they are not constrained to be between 0 and 1 or to sum to 1. However, by exponentiating the redundant variable in (1), we prevent predictive values less than zero, and by dividing by the sum of all the exponentiated variables, we ensure that

$p_{r,g,s,p}^1 + p_{r,g,s,p}^2 + p_{r,g,s,p}^3 + p_{r,g,s,p}^4 + p_{r,g,s,p}^5 = 1$. Using the redundant variables in the hierarchical model makes it easier to specify and fit, without loss of necessary constraints. However, because of the constraint that the probabilities sum to 1, there are only four random variables. To avoid overparameterization we fix $\phi_{r,g,s,p}^5 = 0$.

The hierarchical model is specified for the random redundant parameters:

$$\begin{aligned} \phi_{r,g,s,p}^i &\sim N(\alpha_{s,p}^i, \tau_{s,p}^2) \\ \tau_{s,p}^2 &\sim \text{Gamma}(0.01, 0.01) \\ \alpha_{s,p}^i &\sim N(0, .01) \end{aligned}$$

The predictive values are a function of $\phi_{r,g,s,p}^i$, which are assumed exchangeable across the six combinations of categories of race/ethnicity and gestational weight gain (r,g), but not across self-reported prepregnancy BMI category or term/preterm status (s,p). By separately estimating predictive values by term/preterm we allow for the possibility of differential misclassification. These redundant parameters are shrunk toward grand means, $\alpha_{s,p}^i$, for their self-reported prepregnancy BMI and term/preterm category. This implies that the predictive

values in each of the six combinations of race/ethnicity and gestational weight gain will tend to be shrunk toward one another. The distribution of $\phi_{r,g,s,p}$ includes a precision parameter, $\tau_{s,p}^2$ that is self-reported prepregnancy and term/preterm category-specific. We note that these normal distributions are parameterized in terms of the mean and precision (1/variance). A more complicated covariance structure is possible; however, we found that with our sparse data, specifying a single variance parameter encouraged more shrinkage.

As an example, consider women who have term births ($p=0$) and self-report their prepregnancy BMI to be in the normal weight category ($s=2$). Among these women, there are six combinations of participants (two race/ethnicity x three gestational weight gain categories). For each combination, we want to estimate 5 predictive values:

$p_{r,g,s=2,p=0}^1, p_{r,g,s=2,p=0}^2, p_{r,g,s=2,p=0}^3, p_{r,g,s=2,p=0}^4, p_{r,g,s=2,p=0}^5$, corresponding to the probability of an individual woman's true prepregnancy BMI category in that covariate pattern. When estimating these probabilities for each race and gestational weight gain combination, the hierarchical model allows information to be borrowed across the other combinations of race and gestational weight gain within the stratum of term birth and normal self-reported prepregnancy BMI. That is, the predictive probability of truly being underweight given reporting being normal weight and having a term birth for a given race and gestational weight gain combination, $p_{r,g,s=2,p=0}^1$, borrows information from the predictive probability of being underweight given reporting normal weight and having a term birth in the other five combinations of gestational weight gain and race/ethnicity categories. Each predictive value, loosely speaking, will shrink toward the grand mean of all six of the exchangeable predictive values. We did not allow the model of predictive values to borrow information between preterm and term because if there is an association between prepregnancy BMI and early preterm birth, then one would not expect these predictive values to be the same. Further, we did not allow the model to borrow information from groups of women who self-report different prepregnancy BMI categories because this information is clearly not exchangeable. For instance, the probability of truly being underweight if a woman self-reports being severely obese should be very different than the probability of truly being underweight if a woman self-reports being normal weight.

The hierarchical model of the predictive values was estimated using Markov chain Monte Carlo techniques in JAGS²² and called through R v3.2.3.²³ Markov chains were run for 10,010,000 iterations. The initial 10,000 draws were excluded as a burn-in phase and the remaining iterations were thinned (retaining every 1000th draw) to allow for more rapid post processing of the data. Each iteration of the algorithm produces a random draw from the posterior distribution of each predictive value. Estimating the hierarchical model described above produces 10,000 random samples of each of the 60 sets of the redundant parameters, $\phi_{r,g,s,p}^j$. Substituting these parameters into expression (1) gives 10,000 random samples of the 60 sets of predictive probabilities.

These predictive probabilities were used in a probabilistic bias analysis framework to multiply impute the modeled true prepregnancy BMI category for each woman in the Penn MOMS cohort. We assumed a woman's true prepregnancy BMI category followed a multinomial distribution with probabilities conditional on her race/ethnicity, gestational

weight gain category, self reported prepregnancy BMI category and early preterm status. For each of the 10,000 draws that were retained from the Markov chain Monte Carlo algorithm, each woman's modeled true prepregnancy BMI category was randomly sampled from a multinomial distribution. A logistic model was fit with early preterm birth as the outcome and the imputed true categorical prepregnancy BMI as the main exposure of interest. The estimates of association between prepregnancy BMI and early preterm birth were adjusted for maternal education, age, hospital type, race/ethnicity, insurance type, marital status, urban/rural residence, parity and smoking status. The coefficients and variances for each of the four main effects (severely obese vs normal; obese vs normal; overweight vs normal; underweight vs normal) were retained from each model. To incorporate random sampling error in our final estimate, the four coefficients were drawn from a multivariate normal distribution with mean equal to the coefficients from the logistic model and variance matrix equal to the covariance matrix from the logistic regression. The four coefficients were stored, and this procedure was repeated for each of the 10,000 sets of predictive probabilities produced by the Markov chain Monte Carlo algorithm. The median of the 10,000 estimates equaled the point estimate and the 2.5% and 97.5% percentiles equaled the limits of the 95% simulation interval.

For comparative purposes, we repeated the probabilistic bias analysis described above using two additional models for the predictive values. The first model collapsed all data across race/ethnicity and gestational weight gain categories and estimated a set of predictive values for each combination of prepregnancy BMI category and term/preterm status. As described above, this model corresponds to complete borrowing of information within combinations of race/ethnicity and gestational weight gain. The second model separately estimated the predictive values for all 60 combinations of gestational weight gain, race, reported prepregnancy BMI and term/preterm status. This model corresponds to borrowing no information from any other stratum. We completed the specification of the hierarchical model by assuming the prior distribution on the predictive probabilities had a Dirichlet distribution. All hyperprior values in the Dirichlet distribution had a value of 1, a vague prior distribution that is quickly swamped by the observed validation data. As with the hierarchical analysis described above, we sampled 10,000 values from the posterior distribution of the predictive values for each of the two different models and used those random draws in the probabilistic bias analysis as described above.

Results

Demographic and descriptive characteristics of the Penn MOMS cohort (N=773,625) are given in Table 1. Approximately 51% of the sample has a birth-certificate prepregnancy BMI categorized as normal weight, 23% overweight, 12% obese, 10% severely obese, and 4% underweight. The crude risk of early preterm birth in this cohort was 1.3% among normal weight mothers and highest among underweight (1.9%) and severely obese (2.0%) mothers.

Conventionally adjusted associations between prepregnancy BMI category and early preterm birth are reported in Table 2. Without adjustment for misclassification of prepregnancy BMI status, underweight (OR=1.39, 95% CI=1.28, 1.52) and severely obese women (OR=1.43,

95% CI=1.35, 1.52), relative to normal weight women, had the highest risk of early preterm birth. Obese women had a less pronounced increased risk of early preterm birth (OR=1.21, 95% CI=1.14, 1.29), relative to normal weight women. When adjusting for misclassification using any of the three methods (pooled predictive values, separate predictive values, or hierarchical model) the association between underweight and early preterm birth was attenuated. Adjustments for misclassification had a less pronounced impact on the magnitude of the other three associations.

The widths of the 95% simulation intervals from the misclassification-adjusted models were substantially greater than the widths of the 95% confidence intervals in the model without misclassification adjustment. The decrease in precision is likely the result of incorporating the uncertainty in predictive value estimates from the validation substudy. However, correcting for misclassification, even if the predictive values are assumed known with certainty, could increase the width of the 95% simulation intervals as well. Relatively modest differences in effect sizes were seen between the three models that adjusted for misclassification. Estimates from the collapsed predictive value resulted in greater attenuation of the OR for underweight vs normal weight mothers.

The practical advantage of the hierarchical model for predictive values is shown in Figure 1. Each panel shows the predictive value of truly being underweight, normal weight, overweight, obese or severely obese given that the mother delivered preterm and that the mother's self-reported prepregnancy weight classified her as being underweight. The six sets of estimates within each panel are the predictive value for each of the six exchangeable demographic categories (two race/ethnicity by three gestational weight gain categories). Women who reported being underweight on their birth certificate were most likely to be categorized as underweight by medical record review (upper left panel), except for black women in the highest gestational weight gain category (group 6 in the Figure). For those with discordant categorization on the birth certificate and medical record review, all of the underweight women were classified as normal weight. None of them were categorized as overweight or obese by medical record review (last three panels). In the upper-left panel, the hierarchical model shrunk all results towards the grand mean probability of truly being underweight of 0.65. Predictive probabilities of truly being normal weight were shrunk toward the grand mean of 0.35 (upper-right panel). This shrinkage had modest effect on the point estimate for most of the race/ethnicity by gestational weight gain categories (upper two panels); however, the precision of the predictive values was markedly improved for the hierarchical model estimates. For black women with the highest gestational weight gain (group 6 in the figure) the shrinkage was substantial, with the shrunken estimate of correct classification much closer to the grand mean than the maximum likelihood estimate in both of the upper panels. By borrowing this information, the hierarchical regression estimates that the probability of birth certificate underweight being confirmed by medical record is approximately 60%, as opposed to the 25% observed in the sample of only four women.

Discussion

In this study we implemented a semi-Bayes hierarchical model to allow for more flexible implementation of adjustments for misclassification. This model is flexible enough to extend

to a wide range of validation data structures. We fit a Bayesian hierarchical model for predictive values from a balanced design validation substudy, allowing stratum-specific predictive values to be shrunk toward the grand mean of the predictive probabilities across exchangeable strata. Shrinkage techniques can improve estimation and we provide an example in which less plausible stratum specific estimates were shrunk toward the more plausible group mean. Greenland has previously implemented an approach similar to ours when modeling exposure misclassification of magnetic fields and their effect on childhood leukemia.² We are unaware of it having been implemented in the epidemiologic research literature, however.

The largest impact of these adjustments for misclassification occurred for the association between underweight vs normal weight. This misclassification-adjusted association was attenuated substantially, relative to the association that was not adjusted for misclassification. We note that we did not assume non-differential misclassification nor is bias toward the null necessarily expected even with non-differential misclassification when the misclassified variable has more than two levels. An adjustment for misclassification in this Penn MOMS cohort has previously appeared using a different modeling framework.⁹ The point estimates from the hierarchical semi-Bayes model in this manuscript are similar to the point estimates in the earlier manuscript. However, the estimates of uncertainty obtained by the semi-Bayes hierarchical model are larger than the earlier estimates of uncertainty. The discrepancy is likely due to the fact that in our hierarchical model, the predictive values are assumed to be random variables with associated variances. By using the samples from the posterior distribution of the predictive values in our probabilistic bias analysis, we explicitly incorporate that variance in our final uncertainty intervals. The paper by Lash et al.,⁹ on the other hand, assumed the stratum-specific predictive values were fixed (although the Bernoulli trials they implemented would have allowed for some deviation from the fixed value in each Monte Carlo realization of the data) resulting in more precise uncertainty intervals. In general, incorporating the variance of the bias parameters will result in more plausible intervals.

Typical analytic approaches for quantitative bias analysis assume a simple structure for the bias parameters, typically a simple univariate or bivariate distribution. Probabilistic bias analysis proceeds by repeatedly sampling from that distribution and using those samples to adjust the effect estimate of interest. However, when multiple prior studies inform the bias distribution or, as in our situation, a validation dataset is available, there may be benefits to adding additional model structure to the bias parameter data. This is particularly true when, as in this example, the conventional study sample is very large. In this circumstance, much of the uncertainty arises from the validation data, not from the study sample itself, so careful modeling of the validation data has merit.

Probabilistic bias analysis has a distinctly Bayesian flavor. In many situations, the two approaches are virtually indistinguishable.^{2,16,24} We have opted to fit a Bayesian hierarchical model to the validation data and a frequentist regression to the outcome data. The outcome regression model could, however, also be made explicitly Bayesian by placing prior distributions on all parameters in this model. Data augmentation would be easily implemented in this setting.^{25–27} SAS and Stata also have easy-to-implement Markov chain

Monte Carlo procedures. However, specifying reasonable prior distributions could be difficult if previous research was subject to the same misclassification biases and presented without quantitative bias analysis. In this analysis, the size of the observed data was so substantial that we assume most plausible prior distributions would have little impact on final results.

Our method has several limitations. First, we implemented a hierarchical model to estimate the predictive values. Hierarchical models are well known to reduce mean squared error in many settings. However, because there is no free lunch, these models can achieve a reduction in mean squared error by introducing bias. We are not aware of a literature suggesting that reducing the mean squared error of a bias parameter necessarily decreases the mean squared error of the effect estimate of primary interest: the odds ratio associating prepregnancy BMI and early preterm birth, in our case. Second, we have not taken into account any other source of bias such as uncontrolled confounding or misclassification of other covariates which could potentially influence our results. The approach we have taken in this paper could be extended to a multiple bias modeling framework to account for these biases.²⁸ Third, the validation data are suboptimal in several aspects. Measured prepregnancy weight and height were not available in the medical records as a gold standard for prepregnancy BMI categorization. We relied on an ‘alloyed’ gold standard: self-report prepregnancy weight at a median of 9.3 weeks gestation, which has strong agreement with measured preconception measures in other studies.^{29,30} However, it is unknown whether this good agreement applies in our sample. In addition, all of the validation data were collected at a single hospital in Pittsburgh to optimize cost-efficiency, whereas the cohort included births to women in all of Pennsylvania. Fourth, the results of the proposed method will depend on the validity of the modeling assumptions. Of particular note is the assumption of exchangeability of certain groups (race/ethnicity and gestational weight gain) but not others (preterm/term status). If these assumptions do not hold, model results may be suspect.

Given concerns surrounding possible biases in epidemiologic data, quantitative bias analysis should be undertaken much more often.³ Relatively simple techniques, such as the typical application of probabilistic bias analysis, may be sufficient in many cases. However, when there is detailed information regarding the bias parameters (such as the rich and large validation substudy in this instance), extending the bias parameter model may be advantageous.

Acknowledgments

Sources of funding: This project was supported by NIH grant R21 HD065807 and the Thrasher Research Fund (#9181).

References

1. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003; 14(4):451–458. [PubMed: 12843771]
2. Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). *J R Stat Soc Ser A*. 2005; 168(2):267–306.
3. Lash TL, Fox MP, MacLehose RF, Maldonado G, Mccandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014; 43(6):1969–1985. [PubMed: 25080530]

4. Gustafson P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. 2003
5. Lash T, Fox M, Fink A. Applying Quantitative Bias Analysis to Epidemiologic Data. 2011
6. Chu H, Wang Z, Cole SR, Greenland S. Sensitivity Analysis of Misclassification: A Graphical and a Bayesian Approach. *Ann Epidemiol.* 2006; 16(11):834–841. [PubMed: 16843678]
7. MacLehose RF, Olshan AF, Herring AH, Honein MA, Shaw GM, Romitti PA. Bayesian methods for correcting misclassification: an example from birth defects epidemiology. *Epidemiology.* 2009; 20(1):27–35. [PubMed: 19234399]
8. Gustafson P, Le ND, Saskin R. Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics.* 2001; 57(2):598–609. [PubMed: 11414590]
9. Lash TL, Abrams B, Bodnar LM. Comparison of bias analysis strategies applied to a large data set. *Epidemiology.* 2014; 25(4):576–582. [PubMed: 24815306]
10. Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J Clin Epidemiol.* 1990; 43(9):941–947. [PubMed: 2213082]
11. Witte JS, Greenland S, Haile RW, Bird CL. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology.* 1994; 5(6):612–621. [PubMed: 7841243]
12. Greenland S. Principles of multilevel modelling. *Int J Epidemiol.* 2000; 29(1):158–167. [PubMed: 10750618]
13. Gelman, A., Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press; 2007.
14. MacLehose RF, Dunson DB, Herring AH, Hoppin JA. Bayesian methods for highly correlated exposure data. *Epidemiology.* 2007; 18(2):199–207. [PubMed: 17272963]
15. Witte JS, Greenland S. Simulation study of hierarchical regression. *Stat Med.* 1996; 15(11):1161–1170. [PubMed: 8804145]
16. MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology.* 2012; 23(1):151–158. [PubMed: 22157311]
17. Bodnar LM, Siminerio LL, Himes KP, et al. Maternal obesity and gestational weight gain are risk factors for infant death. *Obesity (Silver Spring).* 2016; 24(2):490–498. [PubMed: 26572932]
18. Bodnar LM, Abrams B, Bertolet M, et al. Validity of Birth Certificate-Derived Maternal Weight Data. *Paediatr Perinat Epidemiol.* 2014; 28(3):203–212. [PubMed: 24673550]
19. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics.* 1999; 55(4):1193–1201. [PubMed: 11315067]
20. World Health Organization. *Obesity: Preventing and Managing the Global Epidemic.* 2000.
21. Gelman A, Bois F, Jiang J. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J Am Stat.* 1996; 91:1400–1412.
22. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proc 3rd Int Work Distrib Stat Comput (DSC 2003);* 2003. p. 20–22.
23. R Core Development Team. *R: A Language and Environment for Statistical Computing.* 2016. <https://www.r-project.org>
24. Steenland K, Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol.* 2004; 160(4):384–392. [PubMed: 15286024]
25. Greenland S, Christensen R. Data augmentation priors for Bayesian and semi-Bayes analyses of conditional-logistic and proportional-hazards regression. *Stat Med.* 2001; 20(16):2421–2428. [PubMed: 11512132]
26. Bedrick EJ, Christensen R, Johnson W. A New Perspective on Priors for Generalized Linear Models. *J Am Stat Assoc.* 1996; 91(436):1450–1460.
27. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol.* 2007; 36(1):195–202. [PubMed: 17329317]
28. Goldstein N, Burstyn I, Welles S. Bayesian approaches to racial disparities in Hiv risk estimation among men who have sex with men. *Epidemiology.* 2017; 26(5):637–644.

29. Lederman SA, Paxton A. Maternal reporting of prepregnancy weight and birth outcome: consistency and completeness compared with the clinical record. *Matern Child Health J.* 1998; 2(2):123–126. [PubMed: 10728268]
30. Phelan S, Phipps MG, Abrams B, Darroch F, Schaffner A, Wing RR. Randomized trial of a behavioral intervention to prevent excessive gestational weight gain: the Fit for Delivery Study. *Am J Clin Nutr.* 2011; 93(4):772–779. [PubMed: 21310836]

Figure 1.

Predictive values estimated from the validation data using hierarchical model (triangle) and maximum likelihood estimation (circle). The horizontal line is the grand mean. Panels display the probability of truly being underweight, normal weight, overweight, obese or severely obese for a woman who delivers preterm and whose birth certificate reports underweight. The x-axis within each panel represents the six possible gestational weight gain by race combinations.

Table 1

Descriptive characteristics of the Penn MOMS cohort (2003–2010), N=773, 625.

	Underweight (BMI<18.5), n(%) Total N=31,774	Normal (BMI: 18.5–24.9 kg/ m ²), n(%) Total N= 396,829	Overweight (BMI: 25–29.9 kg/m ²), n(%) Total N=181,017	Obese (BMI: 30–34.9 kg/m ²), n(%) Total N=90,466	Severely Obese (BMI 35 kg/m ²), n(%) Total N=73,539
Age					
<20	5250 (17)	36,496 (9)	12,714 (7)	5,304 (6)	3,090 (4)
20–29	18,256 (57)	194,493 (49)	91,812 (51)	47,218 (52)	39,536 (54)
30	8,268 (26)	165,840 (42)	76,491 (42)	37,944 (42)	30,913 (42)
Race					
White	28,168 (89)	354,276 (89)	152,418 (84)	74,378 (82)	59,212 (81)
Black	3,606 (11)	42,553 (11)	28,599 (16)	16,088 (18)	14,327 (19)
Gestational Weight Gain					
<20 th percentile	6,157 (19)	75,230 (19)	33,682 (19)	16,937 (19)	13,337 (18)
20 th –80 th percentile	18,998 (60)	239,483 (60)	109,589 (61)	54,590 (60)	45,625 (62)
>80 th percentile	6,619 (21)	82,116 (21)	37,746 (21)	18,939 (21)	14,577 (20)
Education					
< High School	6,859 (22)	52,533 (13)	22,436 (12)	11,102 (12)	8,281 (11)
High School	10,285 (32)	94,792 (24)	49,945 (28)	28,609 (32)	25,643 (35)
Some College	7,747 (24)	101,010 (25)	53,325 (29)	28,923 (32)	25,017 (34)
College Graduate	6,883 (22)	148,494 (37)	55,311 (31)	21,832 (24)	14,598 (20)
Smoking Status					
No	19,422 (61)	302,828 (76)	136,547 (75)	66,785 (74)	54,693 (74)
Yes	12,352 (39)	94,001 (24)	44,470 (25)	23,681 (26)	18,846 (26)
Preterm Birth					
Term	31,175 (98)	391,851 (99)	178,582 (99)	88,973 (98)	72,079 (98)
Early preterm	599 (2)	4,978 (1)	2,435 (1)	1,493 (2)	1,460 (2)

Table 2

Odds Ratios^a and 95% Confidence Intervals for the association between BMI category and early preterm birth, with and without misclassification adjustment

	No Misclassification Adjustment	Adjustment with collapsed predictive values	Adjustment with separate predictive values	Adjustment with hierarchical modeling of predictive values
Underweight vs Normal	1.39 (1.28, 1.52)	1.01 (0.63, 1.59)	1.15 (0.68, 1.93)	1.09 (0.73, 1.49)
Overweight vs Normal	1.01 (0.96, 1.06)	0.89 (0.66, 1.19)	1.06 (0.74, 1.53)	1.02 (0.77, 1.37)
Obese vs Normal	1.21 (1.14, 1.29)	1.24 (0.91, 1.68)	1.18 (0.82, 1.66)	1.23 (0.91, 1.62)
Severely Obese vs Normal	1.43 (1.35, 1.52)	1.38 (1.10, 1.70)	1.30 (0.94, 1.77)	1.48 (1.23, 1.75)

^a Associations are adjusted for maternal education, age, hospital type, race/ethnicity, insurance type, marital status, urban/rural residence, parity and smoking status