Published in final edited form as:

Epidemiology. 2009 November ; 20(6): . doi:10.1097/EDE.0b013e3181b2ff66.

Adjustment for Response Bias Via Two-phase Analysis An Application

Katherine J. Hoggatt^a, Sander Greenland^b, and Beate R. Ritz^b

^aDepartment of Epidemiology, University of Michigan, Ann Arbor, MI

^bDepartment of Epidemiology, University of California, Los Angeles, CA

Abstract

Background—Records-based studies often have limited covariate data, leading some researchers to collect survey data on a subset. Results for survey responders may be biased due to selective nonresponse and will be less precise due to the decreased responder sample size. We use data from a study of air pollution and birth outcomes to illustrate how a 2-phase analysis can yield less biased and more precise results.

Methods—Our phase 1 group was a cohort of Los Angeles births from which we obtained a phase 2 group of survey responders. We compared estimates for the odds ratio (OR) between entire pregnancy carbon monoxide (CO) exposure and low birth weight in the first- and second-phase groups, adjusting only for variables available for both groups.

Results—For CO exposure of 1 part per million or higher, the conventional adjusted ORs and 95% confidence intervals for low birth weight were 1.15 (1.06–1.25) and 1.33 (1.06–1.68) for the phase 1 and 2 groups, suggesting a possible response bias and decreased precision in the latter estimate. We performed 2-phase analyses of the survey responders and found results similar to those for the cohort when we accounted for possible differential response by CO exposure. In our final analysis, we included both birth record and survey variables in a 2-phase model corrected for possible response bias. The results from weighted-, pseudo-, and maximum-likelihood were similar: 1.13 (1.03–1.25); 1.14 (1.01–1.29); and 1.10 (0.97–1.24), respectively.

Conclusion—Our approach provides a means of checking for response bias and adjusting both point and interval estimates to account for differential response.

Two-phase (or 2-stage) study designs balance the need to recruit large samples with the need to collect detailed, high-quality exposure and covariate data, all at a reasonable expense.^{1–3} A basic 2-phase study consists of a first-phase group randomly drawn from a larger known population, or, more often, a first-phase sample of convenience treated as if it were randomly drawn from a hypothetical superpopulation. Certain variables, including the outcome and possibly exposures or covariates, are measured for all members of the phase 1 group. This group is then classified into strata based on phase 1 variables, and a phase 2 sample is drawn from this group with the probability of subject selection varying across strata in a known manner. More detailed covariate or exposure data are then collected on the phase 2 sample.

Despite their potential utility, 2-phase designs and analytic methods are seldom used in practice, perhaps because opportunities for their use remain unrecognized and guiding

examples are rare. In this paper, we use 2-phase methods to analyze nested case-control data from the University of California, Los Angeles Environment and Pregnancy Outcomes Study of air pollution and low birth weight (weight: <2500 g at birth) to improve the efficiency of our estimates and to adjust for selection and response bias related to measured variables.

Previous studies of air pollution and birth outcomes have used vital records as the primary or sole source of covariate data, yielding large study groups with limited available data on potential confounders. This survey was designed to collect additional detailed information on potential confounders, permitting their analytic control. Here, we contrast results from conventional logistic regression analyses of the phase 1 birth cohort and phase 2 survey responders to those obtained using a 2-phase analysis. With systematic survey nonresponse, statistical theory alone cannot demonstrate that a 2-phase analysis is correcting for response bias; to explore this, we validate our 2-phase results against comparable birth cohort results. We then exploit the full power of the 2-phase analysis to obtain nonresponse-adjusted results with additional control for potential confounders measured using both birth records and the EPOS survey instrument.

TWO-PHASE DESIGN AND ANALYSIS: GENERAL

Because the methods we use are found in few textbooks, we first provide a general description of the 2-phase methodology. Initially, we classified the source population (the phase 1 group; here, a cohort of births identified from birth records) by outcome status (low birth weight), yielding N_1 individuals who are cases and N_0 individuals who are noncases. We then classified the cases and noncases based on their values for select first-phase covariates, yielding J first-phase strata, with N_{1j} cases and N_{0j} noncases in stratum j . The 2-phase analytic methods assume that within each of the $2 \times J = 2J$ cells defined by the cross-classification of outcome and phase 1 strata, the n_{ij} individuals in cell ij are a random sample from the total N_{ij} in that cell at phase 1, with the sampling fraction in cell $ij = f_{ij} = n_{ij}/N_{ij}$. We define X as the row vector of covariates to be included in the regression model; X will usually include first-phase covariates used to define sampling strata but may include any combination of first- and second-phase covariates. We modeled the probability of the birth weight outcome Y for given values of the predictors X using the standard logistic regression:

$$\Pr(Y=1|X=x) = \text{expit}(x\beta) \equiv \exp(x\beta)/(1+\exp(x\beta)) \quad (1)$$

In a 2-phase analysis employing inverse-probability-weighted likelihood estimation, the model is fit by maximizing a weighted log-likelihood with the weight for the ij cell proportional to the inverse sampling fraction, N_{ij}/n_{ij} .⁴⁻⁶ An alternative pseudo-likelihood analysis constructs an unweighted log-likelihood but adds an intercept-offset term to the model.⁷ Neither weighted likelihood nor pseudo-likelihood is fully efficient in general,⁸ but an efficient maximum-likelihood analysis can be performed by placing constraints on the pseudo-likelihood maximization.⁹ We implement all 3 methods in S-Plus.⁶ The Splus functions used in this analysis can be found at <http://faculty.washington.edu/norm/software.html>. These methods require that there be at least one case and one noncase at phase 2 in each of the J strata ($n_{ij} > 0$ for all ij), otherwise the inverse-sampling fractions and intercept offsets are undefined. More general missing data methods without this limitation have been developed,¹⁰ but these formulas have limited software implementation.

APPLICATION USING SURVEY DATA

Study Population

Methods for the Environment and Pregnancy Outcomes survey have been described in detail elsewhere.¹¹ Briefly, we identified a phase 1 group from eligible births from 111 Los Angeles (LA) County zip codes in the year 2003 ($n = 58,316$). Data on infant and maternal covariates for the phase 1 birth cohort were obtained from California State birth records. A sample was drawn from phase 1 birth cohort strata defined by the cross-classification of zip code group (we sampled 100% of cases from 24 zip codes near monitoring stations and 30% of cases in 87 other zip codes located near major population centers or major roadways), and the combination of birth month and month in which the birth was reported to the Los Angeles County registrar.

Normal-weight term births were 1:1 matched to births that were low weight, preterm, or both within the phase 1 birth cohort sampling strata, which included 2 levels for zip code area and 48 for birth month/reporting month (96 total). In this application we focused on low weight births as cases, and due to the sampling design there was an over-representation of preterm births in the control series, which we accounted for by including preterm status as a phase 1 stratification variable. Thus by design there were 192 phase 1 strata, of which only 71 had an eligible case or control to select into our phase 2 sample. In theory, all 71 strata may need to be accounted for in the analysis; however, we found that stratifying on birth season alone was sufficient to account for the sampling from the 71 phase 1 strata. (The estimates derived using birth season alone differed by less than 10% relative to those derived using the finest possible stratification on 71 phase 1 strata based on birth/reporting month and zip code.)

We sampled 6374 births for the survey but had only a 40% response rate (2543 responses). We noted differences between responders and nonresponders in the distribution of phase 1 birth record covariates that did not define the original sampling strata (in particular maternal education). Therefore, in our 2-phase analysis we used a post hoc stratification defined by phase 1 birth record variables strongly related to response (maternal education) and variables reflecting the original sampling scheme (birth season and preterm birth) (Table 1). In addition, we used CO exposure as a phase 1 stratification variable to account for unmeasured covariates that were associated with both CO exposure and response.

Exposure Assessment

Ambient CO was measured hourly using nondispersive infrared photometry. From the birth records we obtained maternal residential zip code at the time of delivery and linked this to the nearest “best” monitor, taking into account distance, geography, and wind flow patterns. Using infant date of birth we calculated entire-pregnancy average CO exposure.

Covariates

For the phase 1 birth cohort, we obtained data on maternal age, race, education, parity, and season of birth from the birth records. We excluded 1109 individuals with missing covariate values, leaving 57,207 phase 1 individuals with complete birth record covariate data. For the phase 2 survey responders, we collected data on maternal smoking, alcohol consumption, residence in a house with a smoker, and marital/partner status; there were 2322 individuals with no missing values for birth certificate or survey data.

Statistical Analysis

An ordinary case-control study would proceed by analyzing the phase 2 individuals without regard to the larger cohort from which they were drawn, except, possibly, to account for

variables used in matching controls to cases. However, when there is nonresponse in addition to matched selection, we would have to assume that the complete cases were a random sample from the cohort (ie, individuals who did not respond were missing completely at random from the phase 1 group within strata of the covariates controlled in the analysis). In contrast, a 2-phase analysis could in theory account for possible response bias in addition to the matched selection.

To assess whether our 2-phase analysis would correct for apparent response bias, we fit 2-phase models with phase 1 birth record variables as the only covariates and compared the 2-phase results to those obtained using conventional logistic regression on the entire birth cohort. We reasoned that if the 2-phase analysis were accounting for nonresponse, results from the 2-phase analyses of survey responders should be close to those obtained for the birth cohort. We considered 2 aspects of the 2-phase analysis that could have an impact on our results: the choice of 2-phase estimator (weighted-likelihood, pseudo-likelihood, and maximum-likelihood) and the choice of phase 1 stratification variables. Candidate stratification variables were features of the sampling design (pre-term status, season of birth), variables whose distribution differed between the phase 1 birth cohort and the phase 2 survey responders (maternal education), and the exposure of interest (entire-pregnancy ambient CO exposure).

Based on our comparison of the birth cohort and 2-phase results we chose an estimator and stratification for our final 2-phase analysis. In the final 2-phase analysis, we included both birth certificate and survey variables in the outcome model to obtain estimates that were likely less confounded than the birth cohort results (due to the inclusion of survey covariates, which were putative confounders) and less prone to response and selection bias than estimates from conventional logistic regression on the survey responders alone. In addition, by using information on how the phase 2 group was derived from the phase 1 group, we expected to obtain more precise estimates than would be obtained by analyzing the phase 2 responders in isolation.

RESULTS

For most covariates, the associations in the phase 2 sample (those selected for the nested study) were similar to those in the phase 1 birth cohort (Table 2), although the confidence intervals were wider for the phase 2 sample, reflecting the decreased sample size. The greatest proportionate shift in the OR point estimate was for the season of birth July–September, where the ORs were 1.01 and 1.16 for the birth cohort and survey sample, respectively. In contrast, the point estimates for almost all covariates in the phase 2 responders differed from those in the phase 1 birth cohort; in particular, the OR for CO exposure was 1.33 in the phase 2 responders compared with 1.15 in the phase 1 birth cohort and 1.12 in the phase 2 sample. As with the phase 2 sample, the results in the phase 2 responders were less precise than those for the phase 1 birth cohort.

We next compared point estimates and standard errors from 3 two-phase models fit with only first-phase birth record variables, in which the phase 1 stratification was defined by CO exposure alone (Table 3). We included CO exposure in our outcome model, which is equivalent to including a phase 1 stratum indicator variable and allows direct comparisons of the 3 types of likelihood estimates when using model-robust (“sandwich” or “empirical”) variance estimators. For the pseudo-likelihood and maximum-likelihood estimates, only the coefficients for CO exposure varied from the phase 2 responder results in Table 2. There were some differences in the estimated coefficients for covariates in the weighted-likelihood models, but all 3 estimators yielded similar results for the OR with CO exposure.

We next fit a series of weighted-likelihood models using 4 different definitions for the phase 1 strata (Table 4). We chose the stratification based on 3 considerations: we included variables thought to be associated with exposure and selection by design (season of birth and preterm status); we included variables whose distribution differed between the survey responders and the birth cohort (maternal education); and we included exposure to improve the estimates of our parameter of interest by accounting for unmeasured variables associated with both exposure and response. We included phase 1 covariates (maternal age, race, parity, education, and season of birth) in the model as potential confounders. Although the estimates in all 4 weighted-likelihood models differed from those of the ordinary logistic regression results on the phase 2 responders, the differences were small for the first-phase variables not included in the stratification. For the association of CO with low birth weight, OR estimates and standard errors were similar across all 4 models and were comparable to those obtained for the birth cohort. We also fit a weighted-likelihood model with stratification based on season of birth and preterm status alone (ie, without CO exposure); this yielded an estimated CO association of OR = 1.41 (95% CI = 1.12–1.78), which was close to the result from the conventional logistic regression on phase 2 responders and apparently not corrected for response bias (Table 2).

When the stratum effects are modeled and the fitted model is incorrect (for example, if the stratum effects on the outcome are modeled using a single linear term when in fact the relation is quadratic), the weighted-likelihood estimates will approximate the results we would expect for the first-phase group if we had second-phase covariate data available for everyone, whereas the estimates from the pseudo-likelihood and maximum-likelihood algorithms do not have such a simple interpretation. To explore potential differences due to this type of model misspecification, we compared weighted-likelihood results to those from pseudo-likelihood and maximum-likelihood models that did not include individual stratum indicators (Table 5). Specifically, we fit pseudo-likelihood and maximum-likelihood models with strata jointly defined by CO exposure, season of birth, pre-term status, and education, and we included in the outcome model the main effects for CO exposure, season of birth, and education with no term for preterm status and no product terms among variables. The estimated CO associations were similar for all 3 models, but some results for variables used to define the stratification (eg, season of birth) differed across 2-phase models.

Finally, we fit models (logistic regression, weighted-likelihood, pseudo-likelihood, and maximum-likelihood) for the phase 2 responders using covariates from both the birth records and the surveys. We compared these results to those obtained from a conventional logistic regression analysis of the phase 1 birth cohort with only first-phase birth record variables in the model (Table 6). The 2-phase estimates adjusted for birth record and survey variables were similar to those for the birth cohort adjusted only for birth record variables, suggesting the phase 2 survey variables are not strong confounders; however, the OR from the conventional logistic regression fit on phase 2 responders differed from the 2-phase estimates, indicating that simply modeling phase 2 covariates (as opposed to using a 2-phase analysis) did not correct the apparent response bias.

DISCUSSION

In this application, we compared results from both conventional logistic regression and 2-phase analyses of survey responders to those obtained for the cohort from which the responder group derived. Using models with the same set of covariates, we found that conventional results for the phase 2 survey responders differed from those obtained for the birth cohort, suggesting the survey responder results were biased either due to the stratified sampling or, more likely, differential response. This is consistent with the results from Table 2 showing the differences between responders and the birth cohort to be greater than those

between the phase 2 sample and the birth cohort. Using a 2-phase analysis with stratification defined by CO exposure and one or more first-phase variables we were able to obtain results for the responders that were close to those from the full birth cohort. However, a 2-phase analysis that did not include CO exposure as a stratification variable yielded little adjustment to the association between CO exposure and low birth weight; we believe this is because there were unmeasured variables associated with CO exposure that influence response, and the only way to account for this exposure-differential response was to include CO exposure itself as a stratification variable.

Our analyses using both birth cohort and phase 2 survey variables in the same model were intended to replicate as closely as possible what we could expect if we had survey covariate data on the entire birth cohort. In this case, we had no “correct” estimates as a reference, so our comparisons were contrasts across the 2-phase models using different stratifications and estimators. We obtained similar results regardless of estimator and stratification as long as CO exposure was included as a stratification variable, and these 2-phase results were close to those obtained for the birth cohort using only birth record variables. Taken together, our results are consistent with the survey covariates not being strong confounders (conditional on the birth record variables already included in the model) and with nonresponse being a greater source of bias than uncontrolled confounding. This is perhaps not surprising since putative individual-level risk factors (such as maternal smoking) that are strong predictors of infant low birth weight may be only weakly associated with ambient CO exposure and therefore not a major source of confounding.

Previous studies that have used 2-phase methods have focused on increasing efficiency and accounting for features of the sampling design. Breslow and Holubkov⁵ used the 2-phase maximum-likelihood estimator to account for the different stratum-specific sampling fractions in a study of predictors of perinatal mortality. Consistent with our results, they found that 2-phase analyses with models containing only first-phase variables, but fit on the phase 2 group, yielded estimates close to those obtained from the full phase 1 group. Similarly, Breslow and Chatterjee⁶ used weighted-likelihood and maximum-likelihood analyses in a study of Wilm tumor data and compared models for the entire phase 1 data set with those fitted on the phase 2 data alone. With a saturated model containing all stratum effects, the authors obtained 2-phase results with weighted-likelihood and maximum-likelihood similar to those from the full data set; the results were even closer with finer stratification on phase 1 variables. Our example differs from these analyses, however, in that we attempted to account for possible response bias by adding ad hoc phase 1 stratification variables (eg, maternal education and CO exposure) that were not part of the original sampling design.

We estimated the association between low birth weight and ambient CO exposure and covariates without controlling for gestational age. This choice reflects our hypothesized causal model, which posits that risk factors may affect premature birth via pathways that include impaired fetal growth—for which birth weight is an imperfect measure—as well as other via paths not mediated by fetal growth. Among births at a given gestational age, we would therefore expect to observe an association between predictors of birth at that gestational age and birth weight, even in the absence of a direct effect of those predictors on fetal growth itself. This spurious, noncausal association with birth weight is induced by conditioning on gestational age, as would occur, for example, if the analysis were limited to term births or adjusted for gestational age.

Therefore, although we included preterm status as a phase 1 stratification variable to account for the oversampling of normal-weight preterm infants, we did not include preterm status as a predictor (that is, we did not include preterm status as a variable in X). This model

specification presents no difficulty for the weighted-likelihood estimator, since the weighting applies regardless of the model form.^{6,12} However, the algorithms we used for pseudo-likelihood and maximum-likelihood analysis assume that stratum effects have been modeled with the correct functional form (here, logit-linear); thus to compare the 3 methods directly if the model form is unknown, it is best to fit models with separate stratum indicators, as has been done by other authors.⁶ In a pseudo-likelihood or maximum-likelihood 2-phase analysis with stratum indicators included in the outcome model and no product terms between the stratum indicators and other covariates, only the estimates for the stratum indicators will differ from the estimates derived from the same logistic model fitted using maximum likelihood to the second-phase data alone; estimates for other covariates included in the outcome model are unchanged.⁷ In contrast, in a weighted-likelihood analysis all estimates may differ from those obtained from logistic regression on phase 2 data alone; nonetheless, we would expect to see the greatest change in the estimates for those covariates most strongly related to the stratification variables (including the stratification variables themselves). When the fitted model for the stratum effects is incorrect (eg, when the stratum effects are modeled as logit-linear, when in fact they are not), the weighted-likelihood analysis will approximate the results of fitting the wrong model to the entire phase 1 group, whereas pseudo-likelihood and maximum-likelihood do not.⁹ This feature makes the weighted-likelihood estimator attractive for situations like ours where we wish to include phase 1 variables as stratification factors but do not wish to include them in our outcome model. The efficiency of the weighted-likelihood estimators can be improved using the more general weighting method of Robins et al⁸; unfortunately this method is complex and unavailable in common software packages.

The final standard errors of our estimates are not strictly correct, because the derivation of these formulas assumes that the sampling fractions from the phase 1 strata are fixed by design and hence known. In our example, the sampling fractions depend not only on features of our study design, but also on the numbers of individuals who provided complete data. Our sampling fractions were thus estimated from the data, and the usual 2-phase estimators do not account for this additional uncertainty. If our target parameter of interest is a function only of the odds ratios (eg, a logistic regression coefficient), this additional uncertainty will not affect the resulting standard errors for that estimate⁵; for other parameters (eg, risk differences), however, the estimation of the sampling fractions is not ignorable.⁸

A final caveat is that our work does not address another important source of error in studies of ambient air pollution—exposure measurement error—as this was not the emphasis of our paper. Although it is often claimed that error in the air pollution measurement will result in attenuated estimates, the conditions that guarantee attenuation may not hold in examples such as ours. We therefore cannot state with confidence that our results are underestimates of the true association between CO exposure and low birth weight. Further theoretical work and simulation needs to be done to establish the conditions under which the bias will be predictably toward or away from the null.

In conclusion, 2-phase methods facilitate adjustments that can reduce selection and response bias related to measured variables, in addition to improving the efficiency of the estimates. There is little analytic cost involved in using these methods and, depending on the exposure and covariate data available for the phase 1 group, the benefits may be considerable.

Acknowledgments

Supported by NIEHS grant ES10960-02.

We thank Sarah Burgard for helpful comments.

REFERENCES

1. Rothman, KJ.; Greenland, S.; Lash, TL. Modern Epidemiology. 3rd ed.. Philadelphia: Lippincott; 2008. chapter 8.
2. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*. 1982; 38:1025–1032. [PubMed: 7168792]
3. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol*. 1982; 115:119–128. [PubMed: 7055123]
4. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Stat Med*. 1991; 10:739–747. [PubMed: 2068427]
5. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Stat Med*. 1997; 16:103–116. [PubMed: 9004386]
6. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl Stat*. 1999; 48:457–468.
7. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika*. 1988; 75:11–20.
8. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994; 89:846–866.
9. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J R Stat Soc B*. 1997; 59:447–461.
10. Little, RJA.; Rubin, DB. Statistical Analysis With Missing Data. 2nd ed.. Hoboken, NJ: John Wiley & Sons Inc; 2002.
11. Ritz B, Wilhelm M, Hoggatt KJ, Ghosh JK. Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles. *Am J Epidemiol*. 2007; 166:1045–1052. [PubMed: 17675655]
12. Scott AJ, Wild CJ. Fitting logistic regression models in stratified case-control studies. *Biometrics*. 1991; 47:497–510.

TABLE 1

Phase 1 Birth Record Variables in the Phase 1 Birth Cohort and the Phase 2 Environment and Pregnancy Outcomes Study Survey Responders and Nonresponders

	Birth Cohort (n = 58,316) No. (%)	Survey	
		Responders (n = 2543) No. (%)	Nonresponders (n = 3831) No. (%)
Maternal age (years)			
<25	19,125 (33)	815 (32)	1268 (33)
25–34	29,493 (51)	1319 (52)	1886 (49)
35	9695 (17)	409 (16)	676 (18)
Missing	3 (<1)	0 (0)	0 (0)
Maternal race/ethnicity			
White	9130 (16)	433 (17)	419 (11)
Hispanic	38,893 (67)	1691 (67)	2573 (67)
African-American	4133 (7)	185 (7)	392 (10)
Asian/Pacific Islander	3648 (6)	190 (8)	396 (10)
Other	2237 (4)	32 (1)	32 (<1)
Missing	275 (<1)	13 (1)	18 (<1)
Parity			
Multiparous	35,026 (60)	1524 (60)	2406 (63)
Nulliparous	23,273 (40)	1020 (40)	1422 (31)
Missing	17 (<1)	0 (0)	2 (<1)
Education (years)			
<12	22,289 (38)	880 (35)	1605 (43)
12	35,024 (60)	1621 (64)	2139 (57)
Missing	1003 (2)	43 (2)	87 (2)
Season of birth			
January–March	14,033 (24)	640 (25)	886 (23)
April–June	14,039 (24)	612 (24)	952 (25)
July–September	15,535 (27)	614 (24)	1004 (26)
October–December	14,709 (25)	678 (27)	988 (26)
Mean CO exposure for entire pregnancy (ppm)			
<1	40,270 (69)	1699 (67)	2531 (66)
1	18,046 (31)	845 (33)	1299 (34)
Birthweight			
Normal	55,145 (95)	2012 (79)	2996(78)
Low birth weight	3171 (5)	531 (21)	835 (22)
Gestational age			
Full-term	52,392 (90)	1515 (60)	2286 (60)
Preterm	5924 (10)	1028 (40)	1545 (40)

Missing values for individual variables are excluded from frequency counts and percentages.

TABLE 2

Association of Birth Record Covariates With Low Birth Weight (Variables Entered Simultaneously)

	Birth Cohort OR (95% CI)	Survey	
		Sample OR (95% CI)	Responders OR (95% CI)
Maternal age (years)			
<25 ^a	1.00	1.00	1.00
25–34	0.97 (0.89–1.06)	1.05 (0.90–1.22)	0.78 (0.61–1.00)
35	1.48 (1.33–1.66)	1.43 (1.17–1.73)	1.17 (0.84–1.62)
Maternal race/ethnicity			
White ^a	1.00	1.00	1.00
Hispanic	1.40 (1.23–1.59)	1.29 (1.04–1.59)	1.20 (0.86–1.67)
African–American	3.11 (2.68–3.62)	2.91 (2.25–3.75)	2.41 (1.56–3.72)
Asian/Pacific Islander	1.48 (1.26–1.74)	1.38 (1.05–1.80)	1.52 (0.97–2.36)
Other	1.65 (1.13–2.41)	1.62 (0.89–2.94)	1.13 (0.44–2.89)
Parity			
Multiparous ^a	1.00	1.00	1.00
Nulliparous	1.37 (1.26–1.48)	1.51 (1.32–1.73)	1.34 (1.08–1.68)
Education (years)			
<12 ^a	1.00	1.00	1.00
12	0.86 (0.79–0.93)	0.90 (0.78–1.03)	0.99 (0.78–1.25)
Season of birth			
January–March ^a	1.00	1.00	1.00
April–June	0.98 (0.88–1.09)	0.99 (0.83–1.19)	1.34 (1.00–1.81)
July–September	1.01 (0.91–1.12)	1.16 (0.97–1.39)	1.33 (0.98–1.79)
October–December	1.13 (1.01–1.25)	1.19 (1.00–1.43)	1.52 (1.12–2.05)
Mean CO exposure for entire pregnancy (ppm)			
<1 ^a	1.00	1.00	1.00
1	1.15 (1.06–1.25)	1.12 (0.97–1.29)	1.33 (1.06–1.68)

^aReference category.

TABLE 3

Association of Birth Record Covariates With Low Birth Weight, Derived From 3 Two-phase Analyses of Survey Responders: Weighted-likelihood, Pseudo-likelihood, and Maximum-likelihood

	Weighted-likelihood OR (95% CI)	Pseudo-likelihood OR (95% CI)	Maximum-likelihood OR (95% CI)
Maternal age (years)			
<25 ^a	1.00	1.00	1.00
25–34	0.78 (0.61–1.00)	0.78 (0.61–1.00)	0.78 (0.61–1.00)
35	1.20 (0.86–1.67)	1.17 (0.85–1.62)	1.17 (0.85–1.62)
Maternal race/ethnicity			
White ^a	1.00	1.00	1.00
Hispanic	1.20 (0.86–1.67)	1.20 (0.86–1.62)	1.20 (0.86–1.67)
African-American	2.42 (1.56–3.72)	2.41 (1.57–3.71)	2.41 (1.57–3.71)
Asian/Pacific Islander	1.53 (0.98–2.40)	1.52 (0.97–2.36)	1.52 (0.97–2.36)
Other	1.11 (0.43–2.84)	1.13 (0.45–2.85)	1.13 (0.45–2.85)
Parity			
Multiparous ^a	1.00	1.00	1.00
Nulliparous	1.37 (1.09–1.71)	1.34 (1.08–1.67)	1.34 (1.08–1.67)
Education (years)			
<12 ^a	1.00	1.00	1.00
12	0.98 (0.77–1.25)	0.99 (0.78–1.25)	0.99 (0.78–1.25)
Season of birth			
January–March ^a	1.00	1.00	1.00
April–June	1.36 (1.01–1.84)	1.34 (1.00–1.81)	1.34 (1.00–1.81)
July–September	1.34 (0.99–1.81)	1.32 (0.98–1.79)	1.32 (0.98–1.79)
October–December	1.56 (1.16–2.10)	1.52 (1.13–2.04)	1.52 (1.13–2.04)
Mean CO exposure for entire pregnancy (ppm)			
<1 ^a	1.00	1.00	1.00
1	1.19 (1.06–1.34)	1.20 (1.06–1.35)	1.20 (1.07–1.34)

Phase 1 stratification based on CO exposure.

^aReference category.

TABLE 4

Association of Birth Record Covariates With Low Birth Weight From 4 Weighted-likelihood 2-phase Analyses of Survey Responders, Variables Entered Simultaneously

	Weighted-likelihood Analyses			
	1 OR (95% CI)	2 OR (95% CI)	3 OR (95% CI)	4 OR (95% CI)
Maternal age (years)				
<25 ^a	1.00	1.00	1.00	1.00
25–34	0.79 (0.62–1.01)	0.76 (0.59–0.99)	0.77 (0.60–0.99)	0.76 (0.60–0.98)
35	1.19 (0.86–1.65)	1.29 (0.91–1.82)	1.28 (0.91–1.80)	1.23 (0.88–1.73)
Maternal race/ethnicity				
White ^a	1.00	1.00	1.00	1.00
Hispanic	1.24 (0.89–1.72)	1.27 (0.90–1.79)	1.33 (0.95–1.88)	1.33 (0.94–1.88)
African-American	2.47 (1.59–3.83)	2.66 (1.69–4.20)	2.81 (1.78–4.44)	2.83 (1.78–4.50)
Asian/Pacific Islander	1.57 (1.00–2.46)	1.55 (0.98–2.47)	1.63 (1.02–2.60)	1.60 (1.00–2.57)
Other	1.17 (0.46–3.01)	1.35 (0.50–2.61)	1.34 (0.50–3.60)	1.37 (0.52–3.63)
Parity				
Multiparous ^a	1.00	1.00	1.00	1.00
Nulliparous	1.34 (1.07–1.67)	1.33 (1.06–1.67)	1.31 (1.04–1.65)	1.30 (1.03–1.64)
Education (years)				
<12 ^a	1.00	1.00	1.00	1.00
12	1.01 (0.79–1.28)	0.94 (0.73–1.20)	0.97 (0.76–1.24)	0.89 (0.78–1.03)
Season of birth				
January–March ^a	1.00	1.00	1.00	1.00
April–June	1.00 (0.89–1.12)	1.37 (1.00–1.87)	1.00 (0.88–1.12)	1.00 (0.89–1.13)
July–September	1.01 (0.90–1.13)	1.28 (0.94–1.75)	0.98 (0.87–1.11)	0.99 (0.88–1.11)
October–December	1.16 (1.03–1.31)	1.46 (1.08–1.98)	1.13 (1.01–1.28)	1.14 (1.01–1.29)
Mean CO exposure for entire pregnancy (ppm)				
<1 ^a	1.00	1.00	1.00	1.00
1	1.15 (1.05–1.26)	1.15 (1.02–1.30)	1.14 (1.04–1.26)	1.14 (1.04–1.25)

Two-phase results based on sampling strata defined jointly by the following variables:

Weighted-likelihood 1: CO exposure × season of birth.

Weighted-likelihood 2: CO exposure × preterm status.

Weighted-likelihood 3: CO exposure × season of birth × preterm status.

Weighted-likelihood 4: CO exposure × season of birth × preterm status × maternal education.

^aReference category.

TABLE 5

Association of Birth Record Covariates With Low Birth Weight From Pseudo-likelihood and Maximum-likelihood 2-phase Analyses of Survey Responders

	Pseudo-likelihood OR (95% CI)	Maximum-likelihood OR (95% CI)
Maternal age (years)		
<25 ^a	1.00	1.00
25–34	0.74 (0.55–0.99)	0.71 (0.48–1.05)
35	1.33 (0.89–1.98)	1.55 (0.89–2.73)
Maternal race/ethnicity		
White ^a	1.00	1.00
Hispanic	1.24 (0.84–1.82)	1.34 (0.84–2.15)
African-American	3.21 (1.86–5.54)	4.80 (1.93–11.95)
Asian/Pacific Islander	1.41 (0.84–2.38)	1.47 (0.78–2.78)
Other	1.36 (0.47–3.95)	1.59 (0.38–6.62)
Parity		
Multiparous ^a	1.00	1.00
Nulliparous	1.40 (1.07–1.82)	1.51 (1.05–2.18)
Education		
<12 ^a	1.00	1.00
12	0.95 (0.81–1.13)	0.73 (0.59–0.90)
Season of birth		
January–March ^a	1.00	1.00
April–June	1.12 (0.97–1.29)	0.99 (0.86–1.13)
July–September	0.95 (0.82–1.10)	1.26 (1.08–1.47)
October–December	1.07 (0.92–1.24)	1.18 (1.03–1.36)
Mean CO exposure for entire pregnancy (ppm)		
<1 ^a	1.00	1.00
1	1.16 (1.02–1.30)	1.21 (1.07–1.37)

Two-phase results based on sampling strata defined by CO exposure × season of birth × preterm status × maternal education.

^aReference category.

TABLE 6

Adjusted Odds Ratios for Low Birthweight by Entire-pregnancy Mean CO Exposure: A Comparison of the Full Birth Cohort, Survey Responders Analyzed Via Ordinary Logistic Regression, and Responders Analyzed Via a 2-phase Analysis (Weighted-likelihood, Pseudo-likelihood, and Maximum-likelihood)

	Birth Cohort ^a OR (95% CI)	Survey Responders ^b OR (95% CI)	Weighted-likelihood ^{b,c} OR (95% CI)	Pseudo-likelihood ^{b,c} OR (95% CI)	Maximum-likelihood ^{b,c} OR (95% CI)
Mean CO exposure for entire pregnancy (ppm)					
<1 ^d	1.00	1.00	1.00	1.00	1.00
1	1.15 (1.06–1.25)	1.32 (1.05–1.66)	1.13 (1.03–1.25)	1.14 (1.01–1.29)	1.10 (0.97–1.24)

^aLogistic regression model includes phase 1 birth record variables: maternal age, race, education, parity, season of birth.

^bLogistic regression model includes the above phase 1 variables plus the phase 2 EPOS survey variables cigarette smoking, alcohol consumption during pregnancy, living in a house with a smoker during pregnancy, and partner status.

^cPhase 1 strata defined by CO exposure × season of birth × preterm status × maternal education.

^dReference category.