Published in final edited form as:

Epidemiology. 2014 July ; 25(4): 576–582. doi:10.1097/EDE.0000000000000102.

Comparison of Bias Analysis Strategies Applied to a Large Data Set

Timothy L. Lash^a, Barbara Abrams^b, and Lisa M. Bodnar^{c,d,e}

^aDepartment of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA

^bDivision of Epidemiology, School of Public Health, University of California at Berkeley, Berkeley, California

^cDepartment of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pennsylvania, PA

^dDepartment of Obstetrics, Gynecology, and Reproductive Sciences, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, PA

^eMagee-Womens Research Institute, Pittsburgh, Pennsylvania, PA

Abstract

Background—Epidemiologic data sets continue to grow larger. Probabilistic-bias analyses, which simulate hundreds of thousands of replications of the original data set, may challenge desktop computational resources.

Methods—We implemented a probabilistic-bias analysis to evaluate the direction, magnitude, and uncertainty of the bias arising from misclassification of prepregnancy body mass index when studying its association with early preterm birth in a cohort of 773,625 singleton births. We compared 3 bias analysis strategies: (1) using the full cohort, (2) using a case-cohort design, and (3) weighting records by their frequency in the full cohort.

Results—Underweight and overweight mothers were more likely to deliver early preterm. A validation substudy demonstrated misclassification of prepregnancy body mass index derived from birth certificates. Probabilistic-bias analyses suggested that the association between underweight and early preterm birth was overestimated by the conventional approach, whereas the associations between overweight categories and early preterm birth were underestimated. The 3 bias analyses yielded equivalent results and challenged our typical desktop computing environment. Analyses applied to the full cohort, case cohort, and weighted full cohort required 7.75 days and 4 terabytes, 15.8 hours and 287 gigabytes, and 8.5 hours and 202 gigabytes, respectively.

Copyright © 2014 by Lippincott Williams & Wilkins

Correspondence: Timothy L. Lash, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd. NE, Atlanta, GA 30322. tlash@emory.edu.

Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

Conclusions—Large epidemiologic data sets often include variables that are imperfectly measured, often because data were collected for other purposes. Probabilistic-bias analysis allows quantification of errors but may be difficult in a desktop computing environment. Solutions that allow these analyses in this environment can be achieved without new hardware and within reasonable computational time frames.

With the advent of inexpensive data storage and broadband networking, some epidemiologists have developed research projects that use enormous data sets.¹ These large data sources are often queried to answer questions for which they are not ideally suited.² Probabilistic-bias analysis has been suggested as a tool to assess the direction, magnitude, and uncertainty about a bias acting on a study's result.^{3–6} Probabilistic-bias analysis requires simulations, which can be computationally intensive, often entailing 100,000 or more iterations of a simulation to characterize the bias.⁵

These iterated simulations can be implemented on summarized data (eg, 2×2 tables or several strata of 2×2 tables),⁷ by simulating bias terms directly,^{7–9} or by applying the bias model to each record of the data set to simulate the data that a bias-corrected record might contain.^{9,10} Selection bias and bias from confounding can be readily modeled and simulated by either of the first 2 strategies because the observed association can be factorized into the expected association and an error term representing the bias.⁹ In a selection-bias problem, for example, the observed relative estimate of effect (RR_{obs}) can be written as a product of the true relative estimate of effect (RR_{true}) and a bias factor (Bias):

$$RR_{obs} = Bias \times RR_{true}$$

The bias factor is a function of the selection proportions (P_{D^*,E^*}) in the 4 combinations of exposure ($E+$ and $E-$) and disease ($D+$ and $D-$):

$$Bias = \frac{P_{D+E-} P_{D-E+}}{P_{D+E+} P_{D-E-}}$$

In a probabilistic-bias analysis, distributions can be assigned to each of the selection proportions and sampled by Monte Carlo simulation to obtain a distribution of the bias factor. There is no need for data record manipulation or manipulation of summarized data, and so the size of the data set does not affect the computational intensity of the bias analysis.

Most information-bias problems, and especially most misclassification problems, cannot be factorized. For a dichotomous-exposure misclassification problem:

$$RR_{obs} = \frac{a/(a+c)}{b/(b+d)}$$

where a is the observed number of exposed cases, c is the observed number of exposed noncases, b is the observed number of unexposed cases, and d is the observed number of

unexposed noncases. The true relative effect is a function of these frequencies and the positive and negative predictive values for exposure classification (PPV_D and NPV_D):

$$RR_{true} = \frac{aPPV_{D+} + b(1 - NPV_{D+})}{bNPV_{D+} + a(1 - PPV_{D+})} \bigg/ \frac{aPPV_{D+} + b(1 - NPV_{D+})}{+cPPV_{D-} + d(1 - NPV_{D-})} \bigg/ \frac{bNPV_{D+} + a(1 - PPV_{D+})}{+dNPV_{D-} + c(1 - PPV_{D-})}$$

RR_{obs} cannot be factored from the equation for RR_{true} to obtain an estimate of the bias as a function of the predictive values. This is true for most misclassification problems, with few exceptions.⁷ Monte Carlo simulations must, therefore, operate directly on data. The data may be summarized as a crude 2×2 table (as in the equations) or as strata (including strata as finely divided as single records). With stratification, the computational intensity will depend on the size of the data set, which is a function of the number of records and degree of stratification. When simulations are applied to summarized data such as a 2×2 table, an analyst may lose the ability to adjust for multiple covariates. A record-level simulation of misclassification bias¹⁰ is then an option. However, given data sets of hundreds of thousands of records, and the need for at least 100,000 iterations, the computational intensity required may become a barrier, especially for those working with desktop personal computers.

These problems came to the fore when we sought to implement a probabilistic-bias analysis to evaluate the direction, magnitude, and uncertainty of bias arising from a study of the association between prepregnancy body mass index (BMI) and early preterm birth, adjusted for multiple covariates by logistic regression. Using a desktop personal computer to apply the results from a validation substudy to nearly 800,000 eligible birth records by generating 100,000 simulated data sets of equal size immediately raised the specter of a computational problem so intense as to preclude a probabilistic-bias analysis. We therefore designed and implemented a comparison of several possible analytic solutions and compared them with respect to the achieved results and computational resources required for each.

METHODS

Penn MOMS Cohort

Penn MOMS is a cohort study designed to investigate the risk of poor pregnancy outcomes associated with maternal BMI, gestational weight gain, and race/ethnicity. Participants were non-Hispanic white and non-Hispanic black mothers of singleton infants, with birth certificates in Pennsylvania from 2003 to 2010 ($n = 994,020$). We excluded infants with birth certificates that were missing data on birth weight ($n = 4,348$) or gestational age ($n = 15,670$), mother's prepregnancy height or weight ($n = 57,721$), gestational weight gain ($n = 66,384$), or covariates in the final model ($n = 55,731$). Applying these inclusion and exclusion criteria yielded a cohort of 773,625 singleton births. The birth certificate gathers data on self-reported prepregnancy weight and height on the "Mother's Worksheet," which is administered by an interviewer before the mother leaves the hospital. Mothers were aware of whether they gave birth at term or preterm, which may have affected their reporting of

prepregnancy weight. The prepregnancy BMI in the birth certificate was defined as self-reported prepregnancy weight (kg) divided by self-reported prepregnancy height (m) squared and was categorized according to WHO guidelines.¹¹ Gestational age in the birth certificate is based on the best obstetric estimate. Early preterm birth was defined as delivery at <32 weeks' gestation.

Validation Substudy

From this cohort, we selected 46,059 singleton births recorded at Magee-Womens Hospital of the University of Pittsburgh Medical Center, a teaching hospital and tertiary care center with the highest delivery volume of any Pennsylvania facility. Using a balanced design,¹² we selected up to 30 birth certificates at random from those available in 48 strata: 2 race/ethnicity strata (non-Hispanic white, non-Hispanic black), 2 gestational age strata (term = 37 weeks' gestation, preterm = <37 weeks' gestation), 4 prepregnancy BMI strata (underweight [$<18.5 \text{ kg/m}^2$], normal weight/overweight [$18.5\text{--}29.9 \text{ kg/m}^2$], obese [$30\text{--}34.9 \text{ kg/m}^2$], severely obese [$\geq 35 \text{ kg/m}^2$]), and 3 gestational weight gain strata as previously defined¹³ (<20th percentile, 20–80th percentile, >80th percentile of statewide BMI-strata-specific gestational weight gain distributions). Balanced sampling allowed validation within rare subgroups but precluded direct measurement of sensitivity and specificity.¹⁴ Medical record abstractors completed standardized, computer-aided reviews of the selected medical records, including data on prepregnancy weight and height, as reported by mothers at their first prenatal visit. Using these weight and height reports, the medical record prepregnancy BMI was calculated and categorized.

Probabilistic-Bias Analyses

We computed 60 predictive values ($PV_{j,k,l,m}$), one for each combination of $j = 2$ race/ethnicity strata, $k = 2$ gestational age strata (term or early preterm), $l = 5$ BMI categories (underweight, BMI <18.5 ; normal weight, BMI $18.5\text{--}24.9$; overweight, BMI $25\text{--}29.9$; obese class 1, BMI $30\text{--}34.9$; and obese classes 2 and 3, BMI ≥ 35), and $m = 3$ gestational weight gain strata, all as measured on the birth certificate. We used 5 categories of BMI for the bias analysis, rather than the 4 categories used to design the validation substudy, to correspond to WHO BMI classifications. The second BMI category in the validation study was divided into 2 categories for the bias analysis, and classification boundaries were otherwise the same. Therefore,

$$PV_{j,k,l,m} = \frac{r_{j,k,l,m}}{s_{j,k,l,m}}$$

where $r_{j,k,l,m}$ is the number of medical records that yielded the same BMI category as the birth certificate within race/ethnicity category j , gestational age category k , BMI category l , and gestational weight gain category m , and $s_{j,k,l,m}$ is the number of birth certificates for which medical records were reviewed within race/ethnicity category j , gestational age category k , BMI category l , and gestational weight gain category m .

For each of the 100,000 simulations used for a particular analytic strategy, we selected a value for each $PV_{j,k,l,m}$ by sampling from a binomial distribution with mean $s_{j,k,l,m} PV_{j,k,l,m}$

and variance $s_{j,k,l,m} PV_{j,k,l,m} (1 - PV_{j,k,l,m})$ and dividing by $s_{j,k,l,m}$. For each record within a single simulation, we chose the $PV_{j,k,l,m}$ corresponding to the race/ethnicity, gestational age, birth certificate BMI, and gestational weight gain category, and we sampled the number of records assumed to be correctly classified with regard to BMI category by the birth certificate data (x , the “successes” in the usual binomial lexicon) from a binomial distribution with the probability of success $p = PV_{j,k,l,m}$ and the number of trials n equal to the number of birth certificates represented by the record. The number of birth certificates assumed to be incorrectly classified with regard to BMI category by the birth certificate data ($n - x$, the “failures” in the usual binomial lexicon) were redistributed into the remaining BMI categories following a multinomial distribution inferred from the results of the validation substudy pertaining to the particular race/ethnicity, gestational age, birth certificate BMI category, and gestational weight gain stratum.

For example, 23 of 30 white non-Hispanic mothers who gave birth at 32 weeks’ gestation, whose birth certificate weight and height data placed them into the obese class 1, and whose gestational weight gain was >80th percentile remained in the obese 1 BMI category following medical record review

($PV_j = \text{white non-Hispanic}, k = 32 \text{ weeks}, l = \text{obese 1}, m = >80\text{th percentile} = 23/30 = 0.77$). Of the remaining 7 mothers in this stratum whose medical record data placed them in a different prepregnancy BMI category, 2 were overweight according to the medical record review and 5 were obese classes 2 and 3 according to the medical record review. For the purpose of bias analysis, in a particular iteration,

$PV_j = \text{white non-Hispanic}, k = 32 \text{ weeks}, l = \text{obese 1}, m = >80\text{th percentile}$ would be selected from a binomial with mean 30×0.77 and variance $30 \times 0.77 \times (1 - 0.77)$, and divided by 30. If a particular record in the analytic data set represented 100 white non-Hispanic mothers with term birth (32 weeks), obese 1 prepregnancy BMI category according to the birth certificate, and >80th percentile gestational weight gain, and

$PV_j = \text{white non-Hispanic}, k = 32 \text{ weeks}, l = \text{obese 1}, m = >80\text{th percentile}$ equaled its expectation, then one would expect 77 mothers to remain classified as obese 1 in the simulation. Of the remaining 23, one would expect 6.6 ($23 \times 2/7$) to be reclassified as overweight and the remaining 16.4 to be reclassified as obese 2 and 3, following the multinomial distribution inferred from the validation substudy. Over 100,000 iterations, deviations from this expectation will be multinomially distributed.

Analytic Strategy

We used logistic regression to estimate the association between prepregnancy BMI category and delivering early preterm (<32 weeks gestation), adjusted for race/ethnicity, age category, education category, marital status, hospital type, urban/rural residence, nulliparous or parous, smoking status, and type of health insurance.

We implemented our probabilistic-bias analysis strategy with 3 analytic data sets. The first analytic data set contained records for each mother in the Penn MOMS cohort, restricted to 773,625 in the white non-Hispanic and black non-Hispanic race categories with complete data. The second analytic data set included all mothers who gave early preterm birth, and a 10% sample of the cohort of 773,625, representing a case-cohort design.¹⁵ For these first 2

analytic data sets, each record has n fixed equal to 1. The third analytic data set simultaneously stratified the records by the validation-study-design variables: race/ethnicity, gestational age, prepregnancy BMI category, and gestational weight gain category, and by the categories of all the remaining adjustment covariates. Each record represented a unique combination of these variables and was weighted by the frequency with which that combination appeared among the mothers of restricted Penn MOMS cohort. For this third data set, therefore, n was highly variable. We call this the frequency-weighted full cohort analytic data set.

We measured the computational intensity of 6 analyses: 3 conventional logistic regression analyses applied to each of the 3 analytic data sets, and probabilistic-bias analysis applied to each of the analytic data sets. As described elsewhere,^{5,9,10} all the probabilistic-bias analyses required simulation, output, and logistic regression analysis of 100,000 iterations that applied the bias model described above to the analytic data set. Results of each analysis were output to a second data set. Estimates of association stored in this second data set were ranked, and the median, 2.5% and 97.5%, were reported as estimates of the central tendency and simulation interval after iterative application of the bias model. Measures of computational intensity were the size of the simulation data set (not applicable to the 3 conventional analyses) and the central processing unit (CPU) time required to complete all steps of the analysis. All analyses were completed using PC SAS version 9.3 (Cary, NC) on the same desktop personal computer (Dell OptiPlex 790, Intel Core i7 2600 Processor [3.4 GHz, 8 MB] with 1 TB internal hard drive; Windows 7 Professional operating system). SAS code for the frequency-weighted full cohort analysis is available in the eAppendix (<http://links.lww.com/EDE/A786>).

RESULTS

The Table shows the characteristics of the Penn MOMS cohort ($n = 773,625$) restricted to singleton births by non-Hispanic white and black mothers with complete birth certificate data and of the validation subcohort ($n = 1,212$). The balanced sampling design for the validation subcohort allowed more precise estimation of predictive values in substrata that would have been sparsely populated by validating a simple random sample. For example, preterm births comprised 1.4% of births and 39% of the validation subcohort. Figure 1 shows the cross-classification of “birth certificate prepregnancy BMI” group against “medical record prepregnancy BMI” group within categories of the validation study’s design variables. Prepregnancy BMI category derived from medical record review agreed with BMI category derived from the birth certificate for 969 of 1212 reviews (overall agreement of 80%). The 60 predictive values ranged from 25% to 100%, with two-thirds exceeding 75%. In the central 3 BMI groups, more BMI categories were underestimated by self-reported prepregnancy weight on the birth certificate (80) than were overestimated (63).

Figure 2 summarizes the 6 analyses of the association between prepregnancy BMI category and early preterm birth, using normal weight as the reference category. In each of the conventional analyses (the 3-point estimates and intervals on the left side of each grouping), women whose prepregnancy BMI category according to the birth certificate was underweight (OR = 1.39 [95% confidence interval = 1.28–1.52]), obese class 1 (1.21 [1.14–

1.29]), or obese classes 2 and 3 (1.43 [1.35–1.52]) had higher risk of early preterm delivery than women whose prepregnancy BMI category according to the birth certificate was normal weight (according to the weighted cohort analysis). Point estimates and intervals of the full cohort and weighted full cohort were exactly the same. For the case-cohort analytic data set, the point estimates were slightly different and the confidence intervals slightly wider than the full cohort analyses, a result of sampling of the cohort.

All 3 bias analyses yielded similar results (the 3-point estimates and intervals on the right side of each grouping). Women in the underweight category, after bias analysis, had estimates of association nearer the null than in the conventional analysis (OR = 1.15 [95% simulation interval = 1.02–1.30]). Women in the obese class 1 category (1.27 [1.18–1.36]) and obese class 2 and 3 category (1.52 [1.42–1.61]) after bias analysis had estimates of association further from the null than in the conventional analysis. All bias analysis results had wider confidence intervals than their corresponding conventional analysis, reflecting the additional error incorporated into the estimates of association by modeling the error arising from misclassification of self-reported BMI.

The computational intensity required to complete the 3 bias analyses varied substantially, and all were much greater than that required to complete any of the conventional analyses. The CPU times required to complete the conventional analyses for the full cohort, case cohort, and weighted full cohort were 4.1, 0.39, and 0.1 seconds, respectively. The CPU times required to complete the corresponding probabilistic-bias analyses were 7.75 days, 15.8 hours, and 8.5 hours, respectively. The simulation data created by the probabilistic-bias analyses were 4 terabytes for the full cohort, 287 gigabytes for the case cohort, and 202 gigabytes for the weighted full cohort.

Given the substantially lower CPU time and simulation data size for the weighted full cohort, compared with the other 2 probabilistic-bias analyses, we sought to better characterize the underlying explanation. The analysis included 12 variables with 39,120 possible combinations of values assigned to the variables, of which only 22,417 were observed in the cohort of 773,625 mothers. The frequency with which each observed combination appeared ranged from 0 to 20,598. About 39% of the records had frequency of 1 or 2; for these records, there was little advantage to the weighted cohort over the full cohort. Only 0.3% of records had frequency of 1,000 or more; it is these records for which the weighted analysis resulted in substantial gains in computational efficiency compared with the full cohort analysis.

DISCUSSION

The objective of the Penn MOMS cohort study was to investigate the risk of poor pregnancy outcomes associated with maternal BMI, gestational weight gain, and race/ethnicity in a large and complete population. We recognized that self-reported weight collected as part of the birth certificate interview may be subject to error and that these errors may vary depending on the birth outcome. We therefore designed a balanced design validation study,¹² ultimately comparing prepregnancy weight and derived BMI group from the first prenatal visit to the values obtained from the birth certificate. The validation substudy of the

Penn MOMS project consumed a substantial proportion of the project's resources, but still has limitations. First, we limited the cohort to births with complete birth certificate data and the validation subcohort to medical records with complete validation data. Second, medical records were reviewed at only a single hospital, with the largest number of births annually from among all Pennsylvania hospitals. If errors in reporting prepregnancy weight are different outside this select population, these differences would not be captured by the validation study. Allowing binomial error in the predictive values somewhat accounts for this uncertainty. Third, because of the balanced sampling of births into the 48 strata of the validation sample, we could not directly compute the sensitivity or specificity of the classification of prepregnancy BMI categories derived from the self-reported weight on birth certificates.¹⁴ Without measurements of sensitivity and specificity, we could not assess whether errors were differential by birth outcome.⁹

It is imperative that the results of the validation substudy be put to good use, both to achieve the scientific aims and to justify the resources devoted to it. When the analysis plan began to be implemented, it became apparent that the computational burden of applying earlier probabilistic-bias analysis methods¹⁰ to such a large data set would overwhelm typical desktop computing resources. The first option, which is feasible (and we expect would have been successful), would have been to migrate the analysis to an analytic machine up to the task, such as a networked installation of SAS with distributed computing capability. The second option, which we implemented in part, was to upgrade the desktop computing environment. We ultimately installed a 1-terabyte internal hard drive to accommodate the computing requirements of the probabilistic-bias analysis applied to the full cohort. An external hard drive of this size did not have sufficient data transfer speeds to allow the analysis to be completed in a reasonable amount of time.

We recognized that the information density of the 4-terabyte simulation data set was not great. Approximately the same information could be captured by 2 well-understood strategies: case-cohort analysis and frequency weighting of regression records. We expected the first of these to be less computationally intense, but it was not. In retrospect, the reasons are easy to understand. The case-cohort analysis includes a record for each of the 10,965 cases of preterm delivery and records for 10% of the cohort, totaling 88,328 records, each representing only 1 person. The frequency-weighted full cohort data set included only 22,418 records but represented the entire cohort. Its information density, therefore, was far superior to either the full cohort data set or the case-cohort data set. Had there been more covariates or more categories of the covariates, or had we chosen a lower sampling proportion for the case-cohort analysis, the relative computational efficiencies might have changed.

To our knowledge, application of probabilistic-bias analysis methods to a frequency-weighted regression analysis problem has not been previously reported. It is a hybrid solution. Methods to correct for misclassification problems depicted in 2×2 tables date to the 1950s,¹⁶ and probabilistic-bias analysis tools that extend these methods are available.^{5,10} Although this solution can be applied across strata of covariates, and estimates summarized across the strata, this approach is not an option when one wishes to adjust for multiple covariates. The data become too thin, and they cannot easily yield estimates of confidence

intervals for the summarized estimates.¹⁷ Probabilistic-bias analysis solves both problems by applying the same conceptual correction for misclassification to each data record, allowing the covariate information required for adjustment to travel with the corrected value for the potentially misclassified variable. The iterative simulations allow one to construct simulation intervals that incorporate both the additional uncertainty due to information bias and the sampling error encapsulated by the conventional frequentist confidence intervals.^{3,5,9,10} In many applications, these simulation intervals correspond well with the intervals obtained from a full Bayesian analysis.^{18,19} Application of probabilistic-bias analysis to frequency-weighted data sets combines the computational advantages of applying bias analysis to frequencies rather than to individual records, with the computational advantage of adjusting for multiple covariates by regression rather than stratification. Relatively simple generation of simulation intervals provides an additional benefit over the 2×2 table solution. In this application, the results following bias analysis were similar to those obtained from the conventional result, with small differences in the point estimates and wider intervals. This pattern addresses concerns that the conventional results may have arisen from errors, and especially differential errors, in self-reported prepregnancy body weight.

We considered and rejected 2 additional options for applying probabilistic-bias analysis to this large data set. First, we considered calculating a propensity score²⁰ to capture the confounding information encoded by the covariates, and then carrying only this propensity score into the simulation. This strategy would have reduced the dimensions of the data set somewhat, because the number of covariates would have been reduced from 10 to 1. We believed, though, that the computational intensity was being driven by the number of records, not the number of covariates, so we anticipated that the reduction in computational intensity would be negligible. In other large data set problems, where the number of covariates is much larger than we had available, this solution may yield an advantage in computational efficiency.

Second, we considered calculating a distribution-bias factor to capture the confounding information encoded by the covariates, in accordance with earlier methods.^{7,8} This factor could then be subtracted from the simulation (allowing application of the bias analysis solution to a 5×2 table) and reapplied to the summary estimates obtained from the bias analysis. Probabilistic redistribution of the records (or combinations of records) simulated to have been misclassified required either a record-level approach or a frequency-weighted approach, so there was again little to be gained from this solution. For information bias that involves a dichotomous variable (ie, all misclassified information is reclassified into the other category), we expect this solution would be quite effective, and perhaps most effective, at reducing computational intensity.

As data sets grow larger, the accuracy and granularity of the collected data may sometimes suffer. In pooled data projects, data harmonization presents a new avenue for information bias to affect study results.²¹ New approaches to improve the validity of inferences drawn from the data analyses are needed. Probabilistic-bias analysis provides one strategy to estimate the magnitude, direction, and uncertainty arising from these study imperfections.²² Collection of validation data in substudies is a hallmark of good epidemiologic practice and

should not be avoided or diminished out of concern about the computational challenges one might face when using the validation data. Computers of sufficient computational capacity to overcome the apparent barrier presented by applying simulations to large data sets are available in most research settings. Even when that computing power is not available, or is too expensive to use, alternative strategies can be derived without much difficulty. In our application, the frequency-weighted full cohort analysis yielded exactly the same results as the full cohort analysis, with substantially less computing time and substantially lower data simulation requirements. The solutions we implemented and others we suggest above may compare differently depending on the nature of the computing environment, data set, and planned analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

We thank Sara Parisi and Sarah Pugh for data management support.

Supported by NIH grant R21 HD065807 and the Thrasher Research Fund (9181).

REFERENCES

1. Hernán MA, Savitz DA. From “big epidemiology” to “colossal epidemiology”: when all eggs are in one basket. *Epidemiology*. 2013; 24:344–345. [PubMed: 23549177]
2. Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? *Epidemiology*. 2011; 22:292–294. [PubMed: 21464647]
3. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003; 14:451–458. [PubMed: 12843771]
4. Phillips CV, LaPole LM. Quantifying errors without random sampling. *BMC Med Res Methodol*. 2003; 3:9. [PubMed: 12892568]
5. Lash, TL.; Fox, MP.; Fink, AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer; 2009.
6. Greenland S. Multiple-bias modeling for analysis of observational data. *J R Stat Soc Ser A*. 2005; 168:267–308.
7. Lash TL, Schmidt M, Jensen AØ, Engbjerg MC. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiol Drug Saf*. 2010; 19:638–644. [PubMed: 20535760]
8. Bodnar LM, Tang G, Ness RB, Harger G, Roberts JM. Periconceptional multivitamin use reduces the risk of preeclampsia. *Am J Epidemiol*. 2006; 164:470–477. [PubMed: 16772374]
9. Greenland, S.; Lash, TL. Bias analysis. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. *Modern Epidemiology*. 3rd ed.. Philadelphia, Pa: Lippincott Williams & Wilkins; 2008. p. 345-380.
10. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol*. 2005; 34:1370–1376. [PubMed: 16172102]
11. WHO. Technical Report Series 894: Obesity: Preventing and Managing the Global Epidemic. Geneva, Switzerland: World Health Organization; 2000.
12. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics*. 1999; 55:1193–1201. [PubMed: 11315067]

13. Hutcheon JA, Platt RW, Abrams B, Himes KP, Simhan HN, Bodnar LM. A weight-gain-for-gestational-age z score chart for the assessment of maternal weight gain in pregnancy. *Am J Clin Nutr*. 2013; 97:1062–1067. [PubMed: 23466397]
14. Marshall RJ. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J Clin Epidemiol*. 1990; 43:941–947. [PubMed: 2213082]
15. Rothman, KJ.; Greenland, S.; Lash, TL. Case-control studies. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. *Modern Epidemiology*. 3rd ed.. Philadelphia, Pa: Lippincott, Williams & Wilkins; 2008. p. 111-127.
16. Bross I. Misclassification in 2x2 tables. *Biometrics*. 1954; 10:478–489.
17. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Stat Med*. 1988; 7:745–757. [PubMed: 3043623]
18. 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; 160:384–392. [PubMed: 15286024]
19. MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology*. 2012; 23:151–158. [PubMed: 22157311]
20. Greenland, S. Introduction to regression modeling. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. *Modern Epidemiology*. 3rd ed.. Philadelphia, Pa: Lippincott, Williams & Wilkins; 2008. p. 418-455.
21. Dugas M, Fritz F, Krumm R, Breil B. Automated UMLS-based comparison of medical forms. *PLoS One*. 2013; 8:e67883. [PubMed: 23861827]
22. Maldonado G. Adjusting a relative-risk estimate for study imperfections. *J Epidemiol Community Health*. 2008; 62:655–663. [PubMed: 18559450]

Medical Record Data	Birth Certificate Data													
	BMI Group	Term Birth					Pred. Values	BMI Group	Preterm Birth					Pred. Values
		I	II	III	IV	V			I	II	III	IV	V	
	White Race													
	I II III IV V	<20th % Gest. Wt. Gain					I II III IV V	<20th % Gest. Wt. Gain						
		28	5	0	0	0		85%	10	5	0	0	0	67%
		0	19	4	0	0		83%	0	22	0	0	0	100%
		0	4	4	0	0		50%	0	1	6	0	0	86%
		0	1	7	23	0		74%	0	0	3	11	1	73%
	0	0	2	4	25	81%	0	0	0	3	11	79%		
I II III IV V	20%-80th % Gest. Wt. Gain					I II III IV V	20%-80th % Gest. Wt. Gain							
	21	12	0	0	0		64%	17	6	0	0	0	74%	
	0	22	1	0	0		96%	0	21	2	0	0	91%	
	0	0	7	1	0		88%	0	0	6	0	0	100%	
	0	1	1	26	3		84%	0	0	2	21	6	72%	
0	0	1	2	29	91%	0	0	0	1	27	96%			
I II III IV V	>80th % Gest. Wt. Gain					I II III IV V	>80th % Gest. Wt. Gain							
	25	7	0	0	0		78%	12	8	0	0	0	60%	
	0	20	4	0	0		83%	0	20	4	1	0	80%	
	0	0	5	2	0		71%	0	0	4	0	0	100%	
	0	0	2	23	5		77%	0	0	1	14	5	70%	
0	0	0	1	30	97%	0	0	0	1	24	96%			
Black Race														
I II III IV V	<20th % Gest. Wt. Gain					I II III IV V	<20th % Gest. Wt. Gain							
	19	4	0	0	0		83%	6	1	0	0	0	86%	
	1	23	0	0	0		96%	0	22	1	0	0	96%	
	0	2	6	0	0		75%	0	1	4	0	0	80%	
	0	2	15	11	2		37%	0	1	2	9	0	75%	
0	0	4	1	25	83%	0	0	0	0	3	100%			
I II III IV V	20%-80th % Gest. Wt. Gain					I II III IV V	20%-80th % Gest. Wt. Gain							
	21	8	0	0	0		72%	4	4	0	0	0	50%	
	1	16	0	0	0		94%	0	19	0	0	0	100%	
	0	0	12	2	1		80%	0	0	8	1	0	89%	
	0	0	4	16	9		55%	0	0	4	16	3	70%	
0	1	0	3	29	88%	0	0	0	0	26	100%			
I II III IV V	>80th % Gest. Wt. Gain					I II III IV V	>80th % Gest. Wt. Gain							
	18	11	1	0	0		60%	1	3	0	0	0	25%	
	0	15	5	0	0		75%	0	16	1	0	0	94%	
	0	0	9	1	1		82%	0	1	8	3	0	67%	
	0	1	5	19	6		61%	0	0	0	10	5	67%	
0	0	0	1	29	97%	0	0	0	0	16	100%			

FIGURE 1.

Predictive values (pred. values) and cross-classification of prepregnancy BMI category from the birth certificate against prepregnancy BMI category from the medical record, within categories of race, gestational weight gain (gest. wt. gain), and term or preterm delivery. BMI categories are: BMI I = underweight, BMI <18.5 kg/m²; BMI II = normal weight, BMI 18.5–24.9 kg/m²; BMI III = overweight, BMI 25–29.9 kg/m²; BMI IV = obese class 1, BMI 30–34.9 kg/m²; BMI V = obese classes 2 and 3 BMI ≥35 kg/m².

FIGURE 2.

Estimates and 95% intervals associating prepregnancy body mass index categories with early preterm birth. For each comparison group, estimates from left to right are for the conventional full cohort, conventional case cohort, conventional weighted full cohort, probabilistic-bias analysis full cohort, probabilistic-bias analysis case cohort, and probabilistic-bias analysis weighted full cohort.

TABLE

Frequency and Proportion of Penn MOMS Cohort and Validation Subcohort Within Categories of the Validation Substudy Design Variables

	Penn MOMS Cohort				Validation Subcohort			
	Term		Preterm		Term		Preterm	
	No.	(%)	No.	(%)	No.	(%)	No.	(%)
Prepregnancy body mass index								
Underweight (<18.5 kg/m ²)	31,175	(4)	599	(5)	180	(24)	77	(16)
Normal weight (18.5–24.9 kg/m ²)	391,851	(51)	4,978	(45)	131	(18)	129	(27)
Overweight (25–29.9 kg/m ²)	178,582	(23)	2,435	(22)	57	(8)	43	(9)
Obese class 1 (30–34.9) kg/m ²	88,973	(12)	1,493	(14)	182	(25)	114	(24)
Obese classes 2 and 3 (≥ 35 kg/m ²)	72,079	(9)	1,460	(13)	187	(25)	112	(24)
Gestational weight gain; percentile								
<20th	142,216	(19)	3,127	(29)	241	(33)	123	(26)
20th–80th	463,316	(61)	4,969	(45)	250	(34)	194	(41)
>80th	157,128	(21)	2,869	(26)	246	(33)	158	(33)
Race/ethnicity								
White non-Hispanic	660,714	(87)	7,738	(71)	377	(51)	276	(58)
Black non-Hispanic	101,946	(13)	3,227	(29)	360	(49)	199	(42)