Published in final edited form as:

Ann Epidemiol. 2010 January ; 20(1): 23. doi:10.1016/j.annepidem.2009.09.003.

U.S. Maternally Linked Birth Records May Be Biased for Hispanics and Other Population Groups

JACK K. LEISS, PHD, DENISE GILES, MPH, KRISTIN M. SULLIVAN, MPH, RAHEL MATHEWS, MPH, GLENDA SENTELLE, MSHS, and KAY M. TOMASHEK, MD, MPH

Center for Health Research, Constella Group, Durham, NC (J.K.L., K.M.S., R.M.) and the Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, GA (D.G., G.S., K.T.)

Abstract

Purpose—To advance understanding of linkage error in U.S. maternally linked datasets, and how the error may affect results of studies based on the linked data.

Methods—North Carolina birth and fetal death records for 1988-1997 were maternally linked ($n=1,030,029$). The maternal set probability, defined as the probability that all records assigned to the same maternal set do in fact represent events to the same woman, was used to assess differential maternal linkage error across race/ethnic groups.

Results—Maternal set probabilities were lower for records specifying Asian or Hispanic race/ethnicity, suggesting greater maternal linkage error. The lower probabilities for Hispanics were concentrated in women of Mexican origin who were not born in the United States.

Conclusions—Differential maternal linkage error may be a source of bias in studies using U.S. maternally linked datasets to make comparisons between Hispanics and other groups or among Hispanic subgroups. Methods to quantify and adjust for this potential bias are needed.

Keywords

Bias; epidemiologic; Birth certificates; Epidemiologic methods; Medical record linkage; Validation studies

INTRODUCTION

With the recent emphasis on understanding and preventing the recurrence of adverse birth outcomes (1), improving epidemiologic and statistical methods in perinatal epidemiology (2-4), and taking a life-course approach to maternal and child health (5), investigators are increasingly using maternally linked datasets to study perinatal health (6). These datasets consist of state birth and fetal death records covering several consecutive years. They include a variable (the *maternal set number*; see Appendix A for a glossary of terms that appear in italics) that groups the records into *maternal sets*, each set representing the events (births and fetal deaths) that occurred to a particular woman during those years (7,8). Analysis of the data

© 2009 Elsevier Inc. All rights reserved.

Address correspondence to: Jack K. Leiss, Epidemiology Research Program, Cedar Grove Institute for Sustainable Communities, 6919 Lee Street, Mebane, NC 27302. Tel.: 919-647-4740; fax: 919-563-5290. jackl@mcross.org..

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

is based on the assumption that 1) all of the records in a maternal set are for the same woman and 2) all of her records are in that one maternal set.

Assignment of records to maternal sets is the special feature of maternally linked datasets. It allows the investigator to study successive pregnancies to the same woman (9), which otherwise would not be possible using U.S. vital records data. At least 35 studies using maternally linked datasets from eight U.S. states have been published since development of these datasets began in the mid-1990s. These studies have advanced understanding in several areas of perinatal health, including adverse birth outcomes (6,10-12), maternal behaviors (13), the quality of vital records data (14,15), and the quality of maternally linked data (16,17).

In some European countries, maternally linked data sets have been created using the national registration number to identify records that represent different births to the same woman (18-20). This approach works because the national registration number is a unique identifier that is included in the birth records. In the United States, however, birth and fetal death records do not include a unique identifier. Instead, assignment of records to maternal sets is based on similarities and differences across records in names, birth dates, and other personal identifying information. This process, called *maternal record linkage*, is subject to error (*maternal linkage error*) in that records that do not refer to the same woman may nevertheless be assigned to the same maternal set, and records that do refer to the same woman may be assigned to different maternal sets (17).

Despite the increasing reliance on maternally linked datasets in perinatal health research, very little is known about maternal linkage error and how it affects results of analysis of the linked data. This is due in part to a lack of methods, including quantitative measures, for assessing and evaluating maternal linkage error (21). The purpose of this study was to advance understanding of maternal linkage error and how it affects results of studies based on maternally linked data. Our specific objective was to examine differential maternal linkage error across population subgroups. We first adapted the Fellegi and Sunter (22) method of probabilistic record linkage to obtain a quantitative indicator of potential linkage error. We then compared distributions of this indicator across selected population subgroups. Finally, we discuss the implications of our findings for studies based on maternally linked datasets. This continues a line of research whose overall goal is to develop quantitative techniques for adjusting analysis of maternally linked records for potential bias from maternal linkage error (17).

METHODS

The data for this study were North Carolina birth and fetal death records for 1988–1997. Maternal race/ethnicity subgroups were defined by the data contained in the vital records. Race was listed as white, black, American Indian, or Asian. A separate variable indicated whether the mother was of Hispanic ethnicity. Hispanic ethnicity was subcategorized as Cuban, Puerto Rican, Central/South American, or Mexican. This study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill.

Conceptualizing maternal linkage error

Our approach to maternal linkage error was guided by the purpose behind linking the data, namely, statistical analysis of the linked records. This approach suggests that maternal record linkage should be considered analogous to the data collection phase of a study. In this view, the goal of maternal record linkage should be 1) to produce a sample of maternal sets that is representative of the true maternal sets in the population (as opposed to reproducing the true maternal sets themselves) and 2) to enable computation of statistical measures of the representativeness of the maternal sets that can be incorporated into analysis of the data, as is customary with other types of epidemiologic samples. It follows that the assignment of records

to maternal sets should be based on the underlying probability distributions of the relevant population parameters.

A second notion that follows from considering the purpose of maternally linked data is that maternal linkage error should be conceptualized as misclassification of maternal sets (17). This contrasts with previous approaches that defined maternal linkage error as misclassification of pairs of records (21).

Integrating the above concepts, we used *maternal set probability* as our measure of potential linkage error. Maternal set probability was defined as the probability that all of the records in a set refer to the same woman. It was calculated as the mean of the pairwise match probabilities of all possible pairs of records in the set. This is explained in detail below.

Calculating maternal set probability and selecting maternal sets

The Fellegi and Sunter (22) formulas for probabilistic record linkage satisfy the above criterion of being based on underlying probability distributions. However, for linking multiple records within a single file (as opposed to linking records between files), all of the available applications of the Fellegi and Sunter formulas employ practical adaptations that depart from the underlying probability distributions. We therefore used a linkage method that combined the Fellegi and Sunter formulas with probability sampling. Specifically, we calculated the maternal set probability for all possible maternal sets (with restrictions, as described below), and then we selected from the possible maternal sets based on the maternal set probability.

This process had three stages. First, we calculated match probabilities for pairs of records using the Fellegi and Sunter (22) formulas for probabilistic record linkage (using LinkSolv 4.0, Strategic Matching, Inc., Morrisonville, NY). The Fellegi and Sunter approach is based on the assumption that two records that have the same identifying information are more likely to refer to the same individual if that combination of names, birth date, etc. is rare (relative to the data) than if it is common, and similarly, two records that have differing identifying information are less likely to refer to the same individual if the information is rare than if it is common. To calculate pairwise match probabilities, the identifying variables are compared across each pair of records. Variables that have the same value on the two records contribute positively to the pairwise probability, and variables that have differing values contribute negatively to the pairwise probability, both weighted such that rare values have greater weight.

As stated above, the available linkage software departs from the underlying probability distributions when applying the Fellegi and Sunter formulas to linking records within a single file. Therefore, to obtain the pairwise match probabilities, we divided the records into separate files and calculated the pairwise probabilities across files; pairwise probabilities were undefined for records within the same file. This effectively made records in the same file ineligible to be linked to each other, because the linkages were based on the pairwise probabilities (see below). Accordingly, the assignment of records to the separate files was based on the assumption that no two records for a particular woman would have the same birth order (or fetal death order) *and* substantially different personal identifying information (names and birth dates). This assumption is conservative in that it allows for errors or changes in the data across a woman's events (e.g., change in name after divorce or marriage or two records erroneously indicating the same birth order) except for gross errors that would be unlikely to result in linkage in any case. In practice, records were initially organized into separate files based on birth or fetal death order. Records within each file that had similar identifying information were then separated among the files so that they were eligible to be linked with each other. This approach further implied that a particular record could be correctly linked to at most one other record in any one file. We refer to these files below as linkage files.

Using the above approach, we calculated pairwise match probabilities for all possible pairs of records between all possible pairs of linkage files. This made each record eligible to be linked to each of the other records in all of the other files. Our matching strategy for the pairwise probabilities is given in Appendix B.

The second stage of the linkage process involved creating a universe of all possible maternal sets and calculating their maternal set probabilities. To reduce errors as well as computation time, potential sets whose probabilities were 10% or less were excluded from the universe of possible sets.

The third stage involved selecting maternal sets at random proportional to the maternal set probability from the universe of possible sets. In the universe of possible sets, any one record was contained in numerous sets. In stage 3, when a particular set was selected, all other possible sets containing any of those records became ineligible for subsequent selection. Thus, in the selected sample, any particular record was included in one maternal set only.

Stages 2 and 3 were applied, in sequence, to the linkage files in a stepwise fashion; that is, first two files were processed through stage 2 and then stage 3, then a third file was processed through stage 2 and stage 3, then a fourth file, and so on for all of the files. Once a particular set was selected, those records remained together as a set throughout subsequent iterations, although additional records could be added to the set. In this way, the final maternal sets were built up one file at a time. This process is explained further below and illustrated in Table 1.

In the initial iteration of this process, the universe of possible maternal sets (stage 2) consisted of all pairs of records between two of the linkage files. In this initial iteration only, the pairwise probabilities calculated in stage 1 constituted the maternal set probabilities for the possible maternal sets. Implementing stage 3, some of these pairs of records were selected into the sample of maternal sets for that iteration.

In the second iteration, the universe of possible maternal sets included the sets selected in the previous iteration as well as the individual records from the two initial files that were not included in the selected sets. The universe of possible maternal sets for the second iteration was constructed by combining each of these units (i.e., two-record sets selected in the previous iteration and single, unselected records) with each of the records in a third linkage file. Thus, some of the possible maternal sets in the second iteration contained three records and some contained two records. The maternal set probability was calculated for each possible set as the mean of the pairwise match probabilities (from stage 1) of all possible pairs of records in the set. Implementing stage 3, some of the three-record sets and some of the two-record sets were selected into the sample of maternal sets for that iteration.

The third and subsequent iterations followed the pattern of the second iteration. The universe of possible sets was constructed by combining the selected maternal sets from previous iterations and the single records that were not included in any selected maternal set with the records in another linkage file. For example, the universe of possible sets for the fourth iteration included the selected sample from the third iteration that contained four-record, three-record, and two-record sets; three-record and two-record sets that were selected in the second iteration but not in the third iteration; two-record sets that were selected in the first iteration but not in the second or third iterations; and the single records from all of the previous linkage files that were not included in any of the previously selected maternal sets. Each of these sets and single records brought forward from the third iteration was combined with each of the records from another linkage file to form the universe of possible sets for the fourth iteration. The maternal set probabilities were calculated, and the sample of maternal sets was reselected.

The linkage files were entered into the above process according to their nominal birth order. The final maternally linked dataset consisted of the sets selected in the final iteration and the individual records that were not included in the selected sets. Thus, the final dataset included all of the records from the original ten-year birth and fetal death cohort.

Assessing differential linkage error across population subgroups

To assess differential linkage error, we compared the distributions of maternal set probability among selected population groups. Because the distributions were highly skewed to the right, we based the comparisons on the proportion of sets in each group for which the set probability was ≥ 0.99 . Lower proportions were taken as indicating more linkage error.

We postulated that maternal linkage error was more likely to occur for records representing fetal deaths or teenage mothers (maternal age < 20 years) and for large maternal sets. Consequently, we expected maternal set probabilities to be lower in sets that contained at least one fetal death record, at least one teen birth or teen fetal death record, or five or more records. We also compared set probabilities across race/ethnic groups.

RESULTS

The dataset consisted of 1,021,008 birth records and 9,021 fetal death records from events that occurred in North Carolina in 1988–1997. The linkage produced 241,783 maternal sets of two or more records (Table 2), 80 percent of which consisted of two records only (Table 3). The linked dataset was almost identical with one produced previously by a traditional Fellegi and Sunter (22) probabilistic record linkage (using AutoMatch) conducted on the same data (17); 96% of the maternal sets in the earlier study were reproduced exactly in the present study.

Overall, 97% of sets had a maternal set probability ≥ 0.99 . Set probabilities were somewhat reduced for large sets and for sets that included a fetal death record. However, contrary to our prediction, sets that included a teen birth had only slightly reduced probabilities. Among the major race/ethnic groups that were indicated on the records, set probabilities were markedly lower among sets that included a record indicating Asian or Hispanic race/ethnicity (see Table 4).

To further elucidate the lower probabilities among sets containing Hispanic records, we compared the four Hispanic subgroups indicated on the records. The reduced probabilities were concentrated in sets that contained a record indicating Mexican origin of the mother. Within this group, sets that contained a record indicating non-U.S. birth of the mother had reduced probabilities, whereas sets that contained a record indicating U.S.-birth of the mother did not (see Table 4).

DISCUSSION

Evaluations of maternal linkage error in the literature are sparse. Nitsch et al. (23) reported the number of maternally linked birth records that contradicted maternal histories obtained by questionnaire. Leiss (17) used measures of logical consistency between pairs of linked records as indicators of whether the two records referred to the same women. These approaches are inadequate because they measure maternal linkage error as misclassification of individual records. In contrast, the purpose of maternally linked data is to study the clustering of phenomena within women (1) or to adjust for that clustering when it is not of interest in itself (3). It follows that maternal linkage error should be measured as misclassification of maternal sets. Leiss (17) proposed a method for measuring maternal linkage error in this way; however, it requires a gold standard comparison file, which is rarely available for maternally linked datasets.

Two previous studies assessed maternal linkage error as misclassification of maternal sets. Adams et al. (16) reported the proportion of maternal sets for which the number of records in the set equaled the number of events indicated on the records in the set. This proportion (86% overall) was greater for white race, more education, being married, smaller sets, having the same father listed on all records in the set, and having no fetal deaths or infant deaths in the set. Most of the differences were small. Croft et al. (24) identified maternal sets within which there were discrepancies in the mother's names, birth date, age, time between births, parity, and other data.

In the present method, assignment of records to maternal sets is based on the probability that all of the records in the set refer to the same woman, i.e., the probability that the *set* is correctly classified. Differential distributions of this probability were used as indicators of differential maternal linkage error. We found that the distribution of maternal set probability was shifted to the left for maternal sets that included a record indicating Hispanic ethnicity compared to sets that did not include any Hispanic record. By successively examining constituent subgroups of the group with reduced probabilities, we determined that it was one particular subgroup among Hispanics—Mexican immigrants—for whom maternal set probabilities were lower compared to other Hispanics and to the overall population.

Analogous to other applications of probability theory in epidemiology, maternal set probabilities quantify the degree of certainty that can be ascribed to the composition of the sets. Our results suggest that a high degree of certainty characterizes the maternal sets for the overall population, for women who had an event as a teen, and for whites, blacks, and American Indians. The reduced probabilities for Mexican immigrants indicate greater uncertainty for the composition of these sets. This may reflect greater maternal linkage error for this population. It is possible, however, that it reflects lower quality information for determining the linkages (e.g., inconsistency in the spelling of a woman's name across records for her different births) but that the correct linkages were nevertheless achieved. Further research is needed to disaggregate the contributions of linkage error and correct low quality linkages to reduced maternal set probabilities.

The probabilistic sampling method used in this study is intended to yield a sample of maternal sets that is representative of the true maternal sets in the population. Within this framework, linkage error is properly conceptualized as departure from representativeness. The relation of error as departure from representativeness to error as departure from the true linkages, which is the approach customarily taken in the record linkage literature (16,17), requires further study. The high congruence of the maternal sets produced in the present study with those produced by a traditional Fellegi and Sunter linkage of the same data (17) suggests that the present method produces the true maternal sets as well as a method designed for that purpose.

Implications for perinatal research

Potential bias as indicated by reduced maternal set probabilities is of concern in at least three types of studies (21). One is direct comparisons with the subgroup that has excess error. Several studies, not using maternally linked data, have found differences in perinatal outcomes between U.S.- and non-U.S.-born Mexican women in the United States (25). Use of maternally linked data could potentially advance understanding of these differences (26-28), but differential linkage error could bias the intergroup comparisons.

Secondly, potential bias is a concern when a subgroup with excess linkage error comprises a substantial proportion of a larger group in the analysis. Births to non-U.S.-born Mexican women comprised 80% of all Mexican births and 53% of all Hispanic births in the maternally linked dataset. Several studies, not using maternally linked data, have compared perinatal outcomes among Mexican and other Hispanic subgroups (29,30) and among Hispanics and

other population groups (31,32). Similarly, teen births to non-U.S.-born Mexicans accounted for 60% of Hispanic teen births in the maternally linked dataset, and differences in teen birth outcomes among Hispanics and other population groups is an active area of research (33,34). Again, use of maternally linked data to advance understanding of these issues (26-28) would potentially be biased by differential linkage error.

Finally, maternally linked data are increasingly being used for follow-up studies, including studying recurrence of perinatal outcomes (6,9,10,12,13,35,36). Differences in recurrence by Hispanic ethnicity have not been studied to date, but this is a logical next step in this emerging area of research (37). Differential maternal linkage error would be of particular concern as a source of bias in these types of studies because the maternal set is the unit of analysis.

Generalizability

An important question following from the present study is whether our finding of greater uncertainty for maternal sets of Mexican immigrants is true for the other U.S. maternally linked datasets. The several existing datasets were developed using a variety of methods, all of them different from the method used in the present study. None of the previous methods produced statistics for quantifying maternal linkage error. Nevertheless, it is possible to make tentative generalizations from our findings to the other datasets because of a fundamental similarity shared by all of the methods used to create them. In all of the methods, the composition of the maternal sets is determined by agreement across records on a limited number of variables (7, 8,12,13,17,38,39). Because maternal names and birth date predominate in determining the linkages, the various methods can be expected to produce similar maternal sets. Therefore, in datasets that contain a substantial number of births to Mexican immigrants, bias of the type we found seems likely.

More generally, our results suggest that some population groups may be characterized by elevated maternal linkage error. Because the race/ethnic composition of the maternal population varies by state (40), the particular population groups affected may be different for different maternally linked datasets. It would be useful to determine which other population groups (if any) are likely to have reduced maternal set probabilities, whether this is a general characteristic of immigrant populations, or whether it is limited to Mexican immigrants.

CONCLUSIONS

Maternal record linkage has generally been conducted with the “competing goals of linking as many records as possible and avoiding incorrect linkages” (16, p. 91). The linkage method used in this paper resolves this conflict into the unified goal of creating a study population of maternal sets that represents the birth and fetal death experience of the source population. As with other methods of assembling data for epidemiologic research, the primary goal in creating a maternally linked dataset for research should be a sample in which the error is nondifferential and can be quantified using the underlying probability distributions; avoiding linkage errors should be a secondary goal.

The primary outstanding methodological question for epidemiologic studies using maternally linked data is whether, as our results suggest, there is important differential maternal linkage error across population groups. Future research should focus on developing methods to quantify differential error in existing maternally linked datasets, account for bias from differential error in studies using maternally linked data, and reduce differential error in the linkage process.

Enthusiasm for the potential of maternally linked data to lead to improvements in perinatal health has not been matched by methodological rigor in developing maternally linked datasets.

Our findings support the validity of extant studies that were based on maternally linked data, primarily because those studies did not examine ethnic differences. Looking to a future in which maternally linked data will play an increasingly prominent role in perinatal health research, it will be important to understand weaknesses in the data and how these might be influencing results. This will allow realization of the full potential of this powerful tool for improving the health of mothers and infants (1,27).

Acknowledgments

This research was supported in part by contract GS-10F-0351K from the Centers for Disease Control and Prevention and grants HD35785 from the National Institute of Child Health and Human Development, CA88757 from the National Cancer Institute, and UR6/CCU417428 from the National Center for Health Statistics. The authors thank Shawn Harris for developing the random selection algorithm program, Abenah Vanderpuije for programming assistance, Dr. Patrick Crockett for statistical advice, and Dr. Russell Kirby for comments on an earlier version of the manuscript. Dr. Ming Yin contributed significantly to the development of the linkage method.

APPENDIX A

Definitions of terms

Maternal linkage error

Erroneous assignment of records to *maternal sets* such that records that do not refer to the same woman are assigned to the same set, or records that do refer to the same woman are assigned to different sets.

Maternal record linkage

The process of assigning birth and fetal death records to *maternal sets*. Assignment is based on similarities and differences across records in names, birth dates, and other personal identifying information.

Maternal set number

A variable indicating which maternal set a particular record belongs to. If there is no *maternal linkage error*, then each value of this variable represents events to one woman only and no two values represent events to the same woman.

Maternal set probability

The probability that all of the records in a set are from the same woman. Maternal set probability is calculated as the mean of the Fellegi and Sunter pairwise match probabilities of all possible pairs of records in the set.

Maternal set

A group of birth and fetal death records designated as representing all of the events that occurred to one woman. If there is *maternal linkage error*, then a maternal set may represent events to more than one woman and one woman's events may be divided among two or more maternal sets. The term maternal set is preferred to sibship because if there is *maternal linkage error*, the offspring are not siblings and because siblings who have the same father and different mothers would, in the absence of *maternal linkage error*, be assigned to different maternal sets.

APPENDIX B

Matching strategy for pairwise probabilities

Software: LinkSolv version 4.0. Join variables (blocking variables)—four passes: mother's maiden name; mother's birth date; mother's state of birth combined with mother's first name; father's last name.

TABLE B.1

Specifications for agreement of variables for generating pairwise probabilities

Parameter	Variables							
	Mother's information						Father's information	
	Birth date	First name	Middle name	Maiden name	Address	State of birth	Birth date	Last name
Probability of error	0.001	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Probability of "correct but different"	0.01	0.01	0.025	0.01	0.01	0.01	0.15	0.15
Weighting factor for "agree"	1	1	1	1	0.5	1	0.5	0.5
Weighting factor for "disagree"	1	1	1	1	0	1	1	1

REFERENCES

1. Wilcox AJ. The analysis of recurrence risk as an epidemiological tool. *Paediatr Perinat Epidemiol* 2007;21:4–7. [PubMed: 17593190]
2. Ananth CV, Platt RW, Savitz DA. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. *Ann Epidemiol* 2005;15:293–301. [PubMed: 15780777]
3. Louis GB, Dukic V, Heagerty PJ, Louis TA, Lynch CD, Ryan LM, et al. Analysis of repeated pregnancy outcomes. *Stat Methods Med Res* 2006;15:103–126. [PubMed: 16615652]
4. Savitz DA, Dole N, Herring AH. Methodologic issues in the design and analysis of epidemiologic studies of pregnancy outcome. *Stat Methods Med Res* 2006;15:93–102. [PubMed: 16615651]
5. Lu MC, Halfon N. Racial and ethnic disparities in birth outcomes: a life-course perspective. *Matern Child Health J* 2003;7:13–30. [PubMed: 12710797]
6. Ananth CV, Getahun D, Peltier MR, Salihu HM, Vintzileos AM. Recurrence of spontaneous versus medically indicated preterm birth. *Am J Obstet Gynecol* 2006;195:643–650. [PubMed: 16949395]
7. Adams MM, Wilson HG, Casto DL, Berg CJ, McDermott JM, Gaudino JA, et al. Constructing reproductive histories by linking vital records. *Am J Epidemiol* 1997;145:339–348. [PubMed: 9054238]
8. Herman A, McCarthy B, Bakewell J, Ward R, Mueller B, Maconochie N, et al. Data linkage methods used in maternally-linked birth and infant death surveillance data sets from the United States (Georgia, Missouri, Utah and Washington State), Israel, Norway, Scotland and Western Australia. *Paediatr Perinat Epidemiol* 1997;11:5–22. [PubMed: 9018711]
9. Kistka ZAF, Palomar L, Boslaugh SE, DeBaun MR, DeFranco EA, Muglia LJ. Risk for postterm delivery after previous postterm delivery. *Am J Obstet Gynecol* 2007;196:241.e241–241.e246. [PubMed: 17346537]

10. Adams MM, Elam-Evans LD, Wilson HG, Gilbertz DA. Rates of and factors associated with recurrence of preterm delivery. *JAMA* 2000;283:1591–1596. [PubMed: 10735396]
11. Petrini JR, Callaghan WM, Klebanoff M, Green NS, Lackritz EM, Howse JL, et al. Estimated effect of 17 alpha-hydroxyprogesterone caproate on preterm birth in the United States. *Obstet Gynecol* 2005;105:267–272. [PubMed: 15684150]
12. Van Naarden Braun K, Autry A, Boyle C. A population-based study of the recurrence of developmental disabilities--Metropolitan Atlanta Developmental Disabilities Surveillance Program, 1991-94. *Paediatr Perinat Epidemiol* 2005;19:69–79. [PubMed: 15670112]
13. Kruse L, Denk CE, Feldman-Winter L, Rotondo FM. Longitudinal patterns of breastfeeding initiation. *Matern Child Health J* 2006;10:13–18. [PubMed: 16231108]
14. Adams M. Validity of birth certificate data for the outcome of the previous pregnancy, Georgia, 1980-1995. *Am J Epidemiol* 2001;154:883–888. [PubMed: 11700240]
15. Green DC, Moore JM, Adams MM, Berg CJ, Wilcox LS, McCarthy BJ. Are we underestimating rates of vaginal birth after previous cesarean birth? The validity of delivery methods from birth certificates. *Am J Epidemiol* 1998;147:581–586. [PubMed: 9521185]
16. Adams MM, Berg CJ, McDermott JM, Gaudino JA, Casto DL, Wilson HG, et al. Evaluation of reproductive histories constructed by linking vital records. *Paediatr Perinat Epidemiol* 1997;11:78–92. [PubMed: 9018730]
17. Leiss JK. A new method for measuring misclassification of maternal sets in maternally linked birth records: true and false linkage proportions. *Matern Child Health J* 2007;11:293–300. [PubMed: 17066311]
18. Rasmussen S, Irgens LM, Albrechtsen S, Dalaker K. Predicting preeclampsia in the second pregnancy from low birth weight in the first pregnancy. *Obstet Gynecol* 2000;96:696–700. [PubMed: 11042303]
19. Surkan PJ, Stephansson O, Dickman PW, Cnattingius S. Previous preterm and small-for-gestational-age births and the subsequent risk of stillbirth. *N Engl J Med* 2004;350:777–785. [PubMed: 14973215]
20. Bakketeig LS, Jacobsen G, Skjaerven R, Carneiro IG, Knudsen LB. Low birthweight and mortality: the tendency to repeat low birthweight and its association with early neonatal and infant morbidity and mortality. *Paediatr Perinat Epidemiol* 2006;20:507–511. [PubMed: 17052288]
21. Adams MM, Kirby RS. Measuring the accuracy and completeness of linking certificates for deliveries to the same woman. *Paediatr Perinat Epidemiol* 2007;21:58–62. [PubMed: 17593198]
22. Fellegi I, Sunter A. A theory for record linkage. *J Am Stat Assoc* 1969;64:1183–1210.
23. Nitsch D, Morton S, DeStavola BL, Clark H, Leon DA. How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen Children of the 1950s study. *BMC Med Res Methodol* 2006;6:15. [PubMed: 16553951]
24. Croft ML, Read AW, de Klerk N, Hansen J, Kurinczuk JJ. Population based ascertainment of twins and their siblings, born in Western Australia 1980 to 1992, through the construction and validation of a maternally linked database of siblings. *Twin Res* 2002;5:317–323. [PubMed: 12537852]
25. Collins JW Jr, Papacek E, Schulte NF, Drolet A. Differing postneonatal mortality rates of Mexican-American infants with United-States-born and Mexico-born mothers in Chicago. *Ethn Dis* 2001;11:606–613. [PubMed: 11763285]
26. Basso O. Options and limitations in studies of successive pregnancy outcomes: an overview. *Paediatr Perinat Epidemiol* 2007;21:8–12. [PubMed: 17593191]
27. McCarthy BJ, Berendes HW. Pregnancy outcome investigation for the 21st century. *Paediatr Perinat Epidemiol* 1997;11:2–4. [PubMed: 9018710]
28. Petrini JR. Promoting the linkage and analysis of successive pregnancy outcome files in the United States. *Paediatr Perinat Epidemiol* 2007;21:63–65. [PubMed: 17593199]
29. Leslie JC, Diehl SJ, Galvin SL. A comparison of birth outcomes among US-born and non-US-born Hispanic women in North Carolina. *Matern Child Health J* 2006;10:33–38. [PubMed: 16362236]
30. Zhu M, Druschel C, Lin S. Maternal birthplace and major congenital malformations among New York Hispanics. *Birth Defects Res A Clin Mol Teratol* 2006;76:467–473. [PubMed: 16933210]
31. Leslie JC, Galvin SL, Diehl SJ, Bennett TA, Buescher PA. Infant mortality, low birth weight, and prematurity among Hispanic, white, and African American women in North Carolina. *Am J Obstet Gynecol* 2003;188:1238–1240. [PubMed: 12748490]

32. Salihu HM, Garces IC, Sharma PP, Kristensen S, Ananth CV, Kirby RS. Stillbirth and infant mortality among Hispanic singletons, twins, and triplets in the United States. *Obstet Gynecol* 2005;106:789–796. [PubMed: 16199637]
33. Branum AM. Teen maternal age and very preterm birth of twins. *Matern Child Health J* 2006;10:229–233. [PubMed: 16228693]
34. Schempf AH, Branum AM, Lukacs SL, Schoendorf KC. Maternal age and parity-associated risks of preterm birth: differences by race/ethnicity. *Paediatr Perinat Epidemiol* 2007;21:34–43. [PubMed: 17239177]
35. Krulwich C, Herman A, Yu K, Johnson Y. Does changing paternity contribute to the risk of intrauterine growth retardation? *Paediatr Perinat Epidemiol* 1997;11:41–47. [PubMed: 9018714]
36. Mostello D, Catlin TK, Roman L, Holcomb WL Jr, Leet T. Preeclampsia in the parous woman: who is at risk? *Am J Obstet Gynecol* 2002;187:425–429. [PubMed: 12193937]
37. Kistka ZAF, Palomar L, Lee KA, Boslaugh SE, Wangler MF, Cole FS, et al. Racial disparity in the frequency of recurrence of preterm birth. *Am J Obstet Gynecol* 2007;196:131.e131–131.e136. [PubMed: 17306652]
38. Li DK, Wi S. Changing paternity and the risk of preeclampsia/eclampsia in the subsequent pregnancy. *Am J Epidemiol* 2000;151:57–62. [PubMed: 10625174]
39. Zhu BP, Le T. Effect of interpregnancy interval on infant low birth weight: a retrospective cohort study using the Michigan Maternally Linked Birth Database. *Matern Child Health J* 2003;7:169–178. [PubMed: 14509412]
40. Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Kirmeyer S, et al. Births: final data for 2005. *Natl Vital Stat Rep* 2007;56

Table 1

Illustration of the first three iterations of the maternal record linkage process using hypothetical linkage files containing records A–E, W–Z, L–N, and P–R, respectively.

First Iteration			Second Iteration		Third Iteration	
Universe of possible sets	Input to next iteration	Universe of possible sets	Input to next iteration	Universe of possible sets	Input to next iteration	
A-W	A-Y	A-Y-L	D-Z-L	D-Z-L-P	D-Z-L-Q	
A-X	B-W	B-W-L	E-N	A-Y-P*	A-Y-P	
A-Y*	D-Z	D-Z-L*	A-Y	E-N-P	C-R	
A-Z	C	C-L	B-W	B-W-P	B-W	
B-W*	E	E-L	C	C-P	E-N	
B-X	X	X-L	X	X-P	X	
B-Y		A-Y-M	M	M-P	M	
B-Z		B-W-M		D-Z-L-Q*		
C-W		D-Z-M		A-Y-Q		
C-X		C-M		E-N-Q		
C-Y		E-M		B-W-Q		
C-Z		X-M		C-Q		
D-W		A-Y-N		X-Q		
D-X		B-W-N		M-Q		
D-Y		D-Z-N		D-Z-L-R		
D-Z*		C-N		A-Y-R		
E-W		E-N*		E-N-R		
E-X		X-N		B-W-R		
E-Y				C-R*		
E-Z				X-R		
				M-R		

* Indicates sets that were selected into the sample of maternal sets for that iteration.

TABLE 2

Maternal sets of two or more records, by selected population characteristics, North Carolina birth (n=1,021,008) and fetal death (n=9,021) records, 1988-1997

Characteristic *	Percent of total sets	Number of sets
All sets	100	241,783
Teen birth	25.3	61,175
Fetal death	2.3	5,545
Large set (5+ records)	0.7	1,738
Race/ethnic groups [†]		
White	68.8	166,279
Black	29.4	71,014
American Indian	1.7	4,222
Asian	1.2	2,990
Hispanic [‡]	2.5	6,004
Cuban	0.1	128
Puerto Rican	0.3	724
Central/S. American	0.4	901
Mexican	1.6	3,951
U.S.-born	0.4	893
Non-U.S.-born	1.3	3,104

* At least one record in the set has the characteristic. Groups are not mutually exclusive.

[†] For individual records, white, black, American Indian, and Asian are mutually exclusive categories; any of these records could also be Hispanic.

[‡] For individual records, the subgroups are mutually exclusive.

TABLE 3

Distribution (%) of maternal sets by number of records in the set and selected population characteristics, North Carolina birth (n=1,021,008) and fetal death (n=9,021) records, 1988-1997

Characteristic [*]	No. of records in the set				Records in sets of 2+ records as % of all records
	2	3	4	5+	
All sets	79.6	16.7	3.1	0.7	52.8
Teen birth	67.7	24.5	6.2	1.6	55.3
Fetal death	54.4	32.3	10.1	3.1	65.6
Race/ethnic groups [†]					
White	82.9	14.8	2.0	0.3	51.1
Black	71.5	21.2	5.7	1.7	57.3
American Indian	73.6	20.4	5.1	0.9	59.7
Asian	80.4	15.6	3.1	0.8	42.8
Hispanic [‡]	82.3	14.8	2.4	0.5	37.6
Cuban	78.1	17.2	3.9	0.8	37.5
Puerto Rican	80.8	16.1	2.4	0.7	36.3
Central/S. American	85.9	11.7	2.0	0.4	35.2
Mexican	81.8	15.4	2.3	0.5	38.6
U.S.-born	80.2	16.0	2.9	0.9	35.5
Non-U.S.-born	82.3	15.3	2.2	0.3	39.4

^{*} At least one record in the set has the characteristic. Groups are not mutually exclusive.

[†] For individual records, white, black, American Indian, and Asian are mutually exclusive categories; any of these records could also be Hispanic.

[‡] For individual records, the categories are mutually exclusive.

TABLE 4

Percent of maternal sets* with set probability ≥ 0.99 , by selected characteristics, North Carolina birth (n=1,021,008) and fetal death (n=9,021) records, 1988-1997

Characteristic [†]	%
All sets	97.1
Perinatal factors	
Teen birth	96.5
Fetal death	93.1
Large set (5+ records)	90.0
Race/ethnicity	
White	97.9
Black	96.0
American Indian	95.4
Asian	85.1
Non-Hispanic [‡]	97.8
Hispanic [§]	84.4
Cuban	96.1
Puerto Rican	95.0
Central/S. American	91.7
Mexican	80.0
U.S.-born	96.3
Non-U.S.-born	75.4

* Two or more records.

[†] At least one record in the set has the characteristic; therefore, groups are not mutually exclusive. For individual records, white, black, American Indian, and Asian are mutually exclusive categories; any of these records could also be Hispanic.

[‡] Sets that do not contain an Hispanic record.

[§] At least one record in the set has the characteristic; therefore, subgroups are not mutually exclusive. For individual records, Cuban, Puerto Rican, Central/South American, and Mexican are mutually exclusive categories. Also for individual records, U.S.-born and non-U.S.-born are mutually exclusive categories.