Feasibility of Linking Long-Term Cardiovascular Cohort Data to Offspring Birth Records: The Bogalusa Heart Study

Emily W. Harville[1] · Marni Jacobs[1,2] · Tian Shu[1] · Dorothy Breckner[1] · Maeve Wallace[1,3]

Abstract

Introduction Researchers in perinatal health, as well as other areas, may be interested in linking existing datasets to vital records data when the existence or timing of births is unknown. *Methods* 5914 women who participated in the Bogalusa Heart Study (1973–2009), a long-running study of cardiovascular health in childhood, adolescence, and adulthood, were linked to vital statistics birth data from Louisiana, Mississippi, and Texas (1982–2010). Deterministic and probabilistic linkages based on social security number, race, maternal date of birth, first name, last name, and Soundex codes for name were conducted. Characteristics of the linked and unlinked women were compared using t-tests, Chi square tests, and multiple regression with adjustment for age and year of examinations. *Results* The Louisiana linkage linked 4876 births for 2770 women; Mississippi linked 791 births to 487 women; Texas linked 223 births to 153 women; After removal of duplicates and implausible dates, this left a total of 5922 births to 3260 women. This represents a successful linkage of 55% of all women ever seen in the larger study, and an estimated 65% of all women expected to have given birth. Those linked had more study visits, were more likely to be black, and had statistically lower BMIs than unlinked participants. *Discussion* Linking unrelated study data to vital records data was feasible to a degree. The linked group had a somewhat more favorable health profile and was less mobile than the overall study population.

Keywords Data collection · Vital statistics · Birth certificates

✉ Emily W. Harville
harville@tulane.edu

Marni Jacobs
mjacobs@childrensnational.org

Tian Shu
tshu1@tulane.edu

Dorothy Breckner
dbreckne@tulane.edu

Maeve Wallace
mwallace@tulane.edu

1   Department of Epidemiology, Tulane School of Public Health and Tropical Medicine, 1440 Canal St Ste 2000, New Orleans, LA 70112, USA

2   Division of Biostatistics and Study Methodology, Children's National Health System, Washington, DC 20010, USA

3   Department of Global Community Health and Behavior, Tulane School of Public Health and Tropical Medicine, New Orleans, LA, USA

Abbreviations

BMI   Body mass index
SSN   Social security number

Significance Statement

What is already known on this subject? Previous research birth cohorts have linked study data to vital records data, as have studies in countries with national databases. *What does this study add?* This provides evidence that linkage to vital records data is possible for a cohort based on a non-pregnancy topic, when data were collected over a long period and existence and timing of births is unknown. The linked women differed in small ways from the unlinked, being less mobile and having a slightly better cardiovascular risk profile.

Background

Pregnancy is not separate from health at other periods of life. Complications of pregnancy have been linked to later cardiovascular, metabolic, and cancer outcomes (Bonamy et al. 2011; Catov et al. 2011; Cnattingius et al. 2005; Nilsen et al. 2005). Many studies wish to make the best use of data already collected, and so may wish to add on pregnancy to an existing study of cardiovascular health, diabetes, or cancer, or a clinical trial. While such linkages are feasible in theory and straightforward in Scandinavian countries, in the U.S. and other countries without national registries, they present certain challenges. The number and date of birth of the children may not be known; participants' names are likely to change; participants move from the state where they were enrolled.

We are unaware of other U.S.-based studies that have attempted to link vital records data for offspring information when the timing of the births was unknown and when the births occurred a long time after study participation. The Bogalusa Heart Study (BHS) is a long-running study of cardiovascular health in childhood, adolescence, and adulthood, including data on almost 6000 women. In this paper, we present the results of a linkage between female BHS participants and vital records data.

Methods

The BHS was begun in 1973 by Dr. Gerald Berenson (Berenson 2001). Surveys of the town's schoolchildren were repeated approximately every 2 years through 1994, examining newly enrolled children as well as re-examining those previously enrolled, with reexamination of adults 18–50 between 1997 and 2009. Data collection on cardiovascular health and early aging is ongoing.

Birth Record Data Linkage

Louisiana birth records were available from 1982 to 2009 and fetal death records from 1999 to 2010, inclusive. In 1989, the US Standard Certificates and Reports issued a major revision to the Standard Certificate of Live Births (Freedman et al. 1988); this analysis includes records both from before and after the revision. Linkage of Louisiana birth record data to BHS data was completed using Link-Pro v3.0 (InfoSoft, Inc., Winnipeg, MB) (Jaro 1995; Nitsch et al. 2006; Tromp et al. 2011) Fig. 1 presents the linkage flow graphically. For 1982–1989 records, linkage variables available included maternal last name, soundex code for last name, race, and year of birth. Soundex is a macro available

in LinkPro that converts character variables to phonetic code in order to compensate for common discrepancies in names. Only observations that matched on all four variables were retained for review. This review entailed visual comparison of addresses in order to further confirm a true match.

From 1990 to 2009, a three stage linkage process was used, including deterministic record linkage based on maternal social security number (SSN), and probabilistic linkage when SSN was unavailable.

In Stage I, an exact match of social security number was sought for each woman with a non-missing SSN and was categorized as follows: 1:1 match (women with only one birth), 1:N match (women with multiple births), and unmatched (including truly nulliparous women and women with missing or potentially typo-error SSN in the birth records). All SSN matches (1:1 and 1:N) were considered definite matches.

Stage II was a probabilistic linkage among women who were previously unmatched by SSN and those who were missing SSN in the BHS data. Linkage was based on maternal date of birth (day, month, year), first name, last name, and Soundex codes for first and last names for a total of seven variables. Records that matched on 4 or fewer variables were excluded as non-matches. Records with exact matches on all seven variables were be classified as true-matches. The remaining records (those matched on 5 or 6 of the 7 variables) were reviewed manually and classified as either true-matches or non-matches. Given the panel design of the BHS, many individuals had repeated visits and occasionally different last names or dates of birth were recorded by study staff. The most frequently reported values were used as variables in the linkage macro, and any alternate values were used in manual review of records matching on 5, 6, or 7 of the variables. Manual review also entailed visual comparison address variables not used in the matching strategy.

Finally, Stage III of the linkage repeated Stage II this time using the child's last name (and Soundex code) from the birth record and maternal last name from the BHS dataset to identify any remaining possible matches. The same rules for minimum number of matching variables and manual review for classification of non-matches and true-matches were applied.

Records with exact matches on all seven variables were classified as true-matches; matches on 5 or 6 the remaining records were reviewed manually. This manual review included checking for obvious typos in names, cross-checking race, and where available, comparing addresses. Addresses were used to decide whether or not to retain a match that may have only matched on 5, 6, or 7 of the other variables, but matching addresses was not a requirement for retaining a match as a true match. For the fetal death file (1999–2010), linkage was based on maternal date of birth and name, child's last name, and associated Soundex codes.
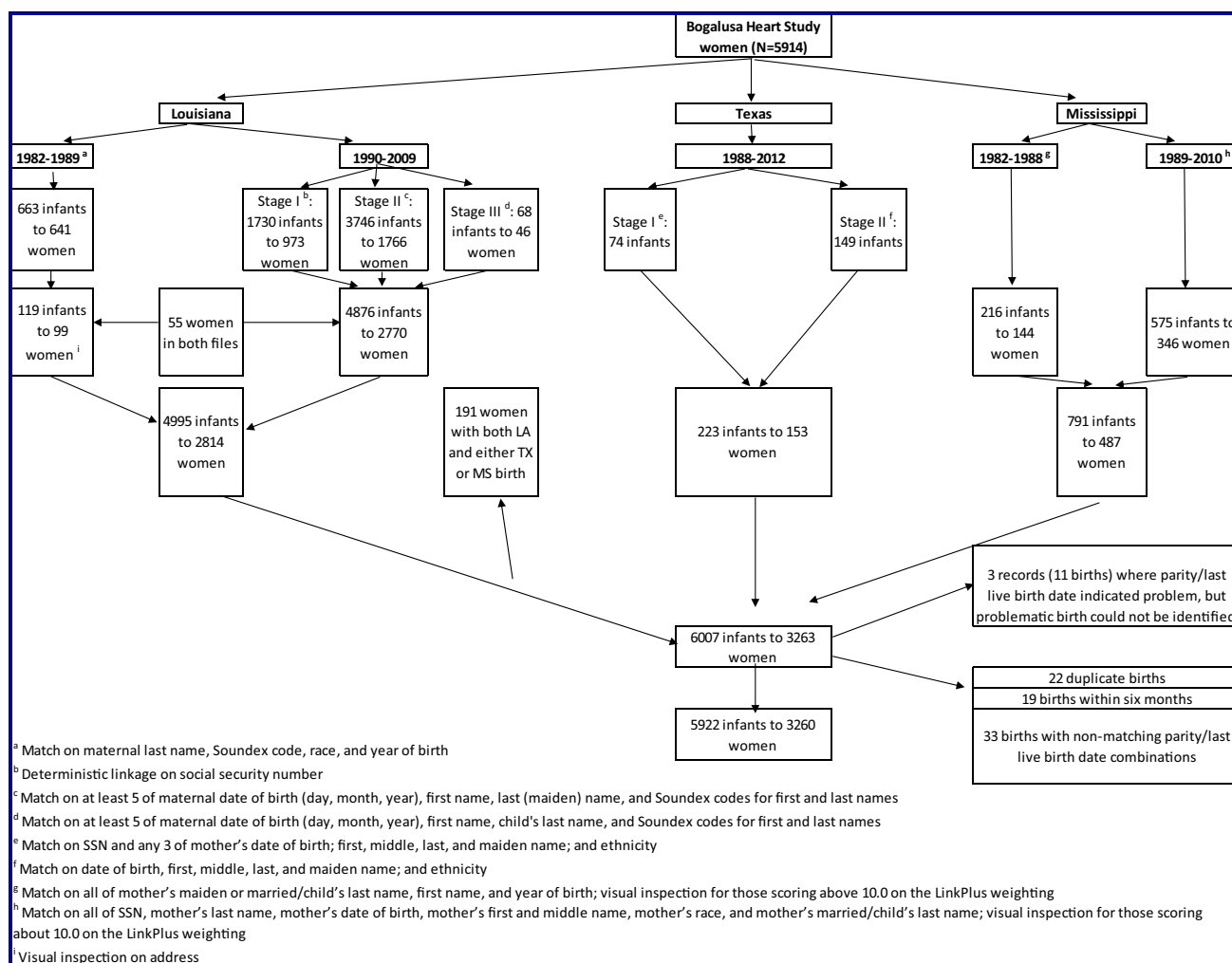
Fig. 1 Flowchart for linkage of Bogalusa Heart Study data to vital statistics data

Slightly different procedures were conducted by the Texas and Mississippi vital statistics departments, based on their internal procedures and policies. Texas conducted a two-stage linkage for data from 1988 to 2012 using Link Plus 3.0 software. The fetal death linkage (from 1991) was based on date of birth, mother's first, middle, and last name. Mississippi also used Link Plus 3.0 (Division of Cancer Prevention and Control 2015), and a blocking and matching procedure. Blocking creates all possible matches between the two files for each pair of blocking variables (last name, SSN, year of birth), then scores the matches, with different weighting for different variables. Mississippi did not link fetal deaths.

Results were then examined for duplicates. If a birth was duplicated or occurred within six months of a previous birth, it was removed from the dataset.

The Institutional Review Boards (IRB) of Tulane University (IRB ID#256406), the State Department of Health and Hospitals of Louisiana, and the Texas Department of State Health Services approved this protocol (Mississippi deferred to the Tulane IRB). The linkage was conducted under a waiver of consent, as it was deemed minimal risk and infeasible without the waiver.

Analysis

The characteristics of those women who were linked and those who were not were compared using Chi square tests for categorical and t-tests for continuous variables. Multiple linear and logistic regression were used to examine whether differences in cardiovascular risk factors between the linked and unlinked groups were explained by the ages and years of the examinations.

Results

A total of 3260 women and 5922 births were linked (Fig. 1), representing a successful linkage of 55% of all women ever seen by BHS examiners. The Louisiana fetal death linkage matched four women to four fetal deaths, and two fetal deaths were identified in the Texas data. Fifteen women were linked to birth records, but had no corresponding data in the BHS database, suggesting that either they had consented to participate at some point but never actually had any data collected, or that they participated only in a small ancillary study. Comparison of those linked and those not indicates that those linked had more study visits, had a younger average age at first visit, had a later average year at most recent visit, and were less likely to smoke. A greater proportion of the linked women were black. (Table 1). They had statistically lower mean BMIs and blood pressure (Table 2). Differences in smoking, cholesterol, and blood pressure were largely explained by differences in the age and calendar year of the examinations of the two groups. BMI continued to be lower in the linked group after adjustment for these factors; the difference was largest for adolescent BMI (-0.59 kg/m^2, $p < 0.01$), and less precise for adult BMI (-0.46 kg/m^2, $p = 0.16$). These patterns basically held when the data were limited to those with 2 or more visits or at least one adult visit (Supplementary Material). For 29% of the women included, the lowest-parity birth included was not their first birth (Table 3).

3413 of these women had information on father's education and 3698 had information on mother's education. 61% of those whose father had less than a high school education were linked, vs. 59% of those whose father had more than a high school education, $p = 0.03$ (percentages for mothers were 61 and 58%, respectively, $p = 0.04$). However, when this difference was adjusted for age and year of visit, or limited to those with adult or multiple visits, it disappeared. Among those whose adult education was known (n = 774, only measured at visits after 2001), higher education was associated with an increased likelihood of linkage (65 vs. 52%, $p < 0.01$), which was not accounted for by age, year, or number of visits.

Discussion

This study demonstrates the feasibility, under some conditions, of linking unrelated study data to vital records data. Based on national data, we would expect approximately 85% of the study participants to have given birth to at least one child, which suggests that this linkage captured at least one

Table 1 Comparison of women linked to vital records data on at least one birth and those not linked

	Linked (n = 3245)		Not linked (n = 2669)		p for difference
	N	%	N	%	
Number of visits					< 0.01
1	949	29.2	1106	41.5	
2	653	20.1	415	15.6	
3–4	798	24.6	526	19.7	
5–6	412	12.7	292	10.9	
7 or more	433	13.3	329	12.3	
Race					0.07
Black	1209	37.3	934	35.0	
White	2036	62.7	1735	65.0	
Ever smoker					0.05
Yes	863	34.9	725	37.8	
No	1612	65.1	1195	62.2	
	Mean	SD	Mean	SD	
Age at youngest visit	8.8	3.5	10.5	6.3	< 0.01
Age at oldest visit	18.8	11.0	18.8	12.3	0.93
Age in 2017	44.2	7.4	46.4	9.3	< 0.01
	Median	p25–p75	Median	p25–p75	
Year at first visit	1981	1974–1987	1978	1974–1988	< 0.01
Year at most recent visit	1993	1986–1996	1989	1982–1994	< 0.01

Table 2 Models of differences in cardiovascular risk according to linkage

	Unadjusted			Adjusted for age, age at first visit, and calendar year		
	Beta	CI	p	Beta	CI	p
BMI	− 0.38	− 0.63, − 0.12	< 0.01	− 0.15	− 0.37, 0.07	0.17
Child BMI	− 0.19	− 0.38, − 0.01	0.04	− 0.13	− 0.31, 0.05	0.16
Adolescent BMI	− 0.08	− 0.39, 0.22	0.59	− 0.59	− 0.90, − 0.27	< 0.01
Adult BMI	0.43	− 0.18, 1.04	0.17	− 0.46	− 1.08, 0.17	0.16
Cholesterol	− 0.14	− 1.56, 1.29	0.85	− 0.80	− 2.22, 0.62	0.27
Child cholesterol	− 0.01	− 1.64, 1.62	0.99	0.02	− 1.60, 1.65	0.98
Adolescent cholesterol	2.33	0.50, 4.16	0.01	− 0.20	− 2.11− 1.71	0.84
Adult cholesterol	− 1.39	− 4.30, 1.52	0.35	− 0.43	− 3.51, 2.64	0.78
Systolic blood pressure	− 1.03	− 1.52, − 0.54	< 0.01	0.03	− 0.40, 0.46	0.91
Child SBP	− 0.02	− 0.53, 0.48	0.93	0.21	− 0.29, 0.70	0.42
Adolescent SBP	− 0.45	− 1.00, 0.10	0.11	− 0.16	− 0.73, 0.42	0.57
Adult SBP	− 1.25	− 2.11, − 0.40	< 0.01	− 0.07	− 0.97, 0.82	0.87

birth for 65% of those who are likely to have given birth. A few previous studies have attempted something similar. The Pregnancy, Infection, and Nutrition study linked to 87% of participants (Vinikoor et al. 2010), within a short time of birth, in a known state of birth. The Children of the 1950s Study (Scotland) was able to link 66% of participants, although their national registry began late enough that it was expected that earlier births would be missed (Nitsch et al. 2006). The Ag Health study identified 94% of reported births and an equal number of unreported livebirths (Romitti et al. 2010). Also, various studies have been able to link birth records to each other or to large administrative databases (Adams et al. 1997a, b; Bell et al. 1994; Emanuel et al. 1992; Herrchen et al. 1997). This linkage differs from these in that the data were collected over a long period of time with the goal of understanding cardiovascular health in children and over the life course, with no idea that pregnancies would be of interest, and so is relevant to studies beyond pregnancy cohorts.

Linkages can produce both false positives and false negatives, and a degree of subjectivity is always present in declaring a match. In 3 cases in the Texas data, for instance, SSN matched but no other variable did, and the values of the other variables were very different. These were judged to be non-matches. Since this study is based in a small area and the first digits of SSN are assigned geographically (Social Security Administration), errors in SSNs may more easily create false positives in this study than in one including births from a broader geographic area. It is likely that many of the initial matches in the early data (4 variables) that could not be confirmed by address were indeed matches, as the address data was often from a different time period. However, given to the possibility of sisters or cousins living in the same area (or simply common names), we preferred to be conservative. Each state applied slightly different

procedures, although the overall set of matching variables and general linkage philosophy were consistent. In ongoing data collection, 81% of participants are from Louisiana, consistent with the proportion in this analysis, arguing for relative consistency in the linkage procedures. However, we have no way of judging the relative quality of linkage across the states within these data. Similarly, we cannot guarantee that linkages that match, particularly those without a SSN, were correct, as we have no gold standard to compare.

Future studies attempting similar linkages will want to consider several issues. Broadly speaking, more data produces a greater likelihood of linkage. Many more records were linked in the post-1990 data, because more information and more specific identifiers were available in the birth records. Similarly, women who participated in later waves were more likely to be linked, in part because SSN was recorded only at later visits, and in part because name changes were more likely to have been reported, leading to a greater chance of matching on mother's or child's last name. Births outside the 1982–2009 time period are possible, and are excluded from the data, as are births from outside the three included states. We cannot distinguish those who were unlinked due never to having borne children from those who were unlinked due to having moved from the area or poor or inconsistent recording of data used for linkage. It is clear, however, that the strongest predictors of linkage were those that indicate data quality or quantity (such as year and number of visits), rather than those that predict childbearing, such as education. The linked women are not a predictably high- or low-risk sample, either medically or socioeconomically; while women with higher education are more mobile, they are also easier to locate if they do not move. While a large proportion of parous women were identified in this analysis, not all of their infants were included, which limits the possibility of

Table 3 Description of children and pregnancies linked to women in the Bogalusa Heart Study (n = 5916)

	N	%
Year of birth		
1982–1989	344	2.8
1990–1994	1397	23.6
1995–1999	1578	26.7
2000–2004	1420	24.0
2005–2009	1116	18.9
2010+	61	1.0
Multiple gestation	183	3.1
State of residence		
Louisiana	4932	83.4
Mississippi	764	12.9
Texas	220	3.7
Interpregnancy interval (among those with prior pregnancy)		
≤ 1 year	275	7.7
> 1–2 years	1055	29.4
2–5 years	1483	41.3
> 5 years	775	21.6
Previous parity at first linked pregnancy		
0	2311	71.2
1	620	19.1
2	226	7.0
3+	87	2.7
Previous gravidity at first linked pregnancy		
0	1928	64.0
1	642	21.3
2	283	9.4
3+	162	5.4
Previous parity at last linked pregnancy		
0	896	27.6
1	1180	36.3
2	744	22.9
3+	429	13.2
Previous gravidity at last linked pregnancy		
0	757	24.3
1	975	31.3
2	744	23.9
3+	635	20.4
Number of linked children per woman (includes multiple births)		
1	1587	48.7
2	1024	31.4
3+	648	19.9
	Mean	SD
Birthweight	3174	613.9
Gestational age	38.4	2.5

focusing on first pregnancy, a common analysis choice. A low number of fetal deaths were linked, both due to the relative rarity of late fetal deaths and, likely, to their under-reporting (Martin and Hoyert 2002). Linkages are probably not a useful method of studying this topic.

Another concern is whether the linked group is representative of the overall study and the population under consideration. With respect to cardiovascular health, the included women have a somewhat more favorable risk profile (lower BMI and blood pressure) than those not included, although absolute differences were not large. As higher cardiovascular risk may be associated with lower fertility (Jacobs et al. 2016)(Wang et al., under review), it is likely that part of this difference is due to unlinked women who genuinely did not have any births. In other cases, this may lead to a less generalizable population for analysis, and perhaps a lower-risk population and lower-power study design. Particularly high-risk women may be under-represented, possibly leading to an underestimate of the effect estimate (if those with worse cardiovascular health and higher risk of complications were preferentially omitted). Second, the less mobile the population and more recent the births, the greater chance of linkage. If mobility is correlated with the outcome under study (or, more likely, with a related socioeconomic factor), then the linked data is likely to be biased. Finally, any linked data are only as good as those included in the birth certificate, which varies by the datapoint being examined—birthweight and gestational age are generally well-recorded, while pregnancy complications like anemia may not be (Vinikoor et al. 2010).

Increasingly, maternal and child health is being related to health in fields other than reproductive health. This analysis indicates the possibility of combining data collected for other purposes with routinely-collected data, for a better understanding of life course influences on pregnancy. In future research, we will explore the correlations between reproductive histories collected through interviews and the vital statistics data, as well as the relative feasibility of the two types of data collection. The extensive longitudinal exposure information collected by the BHS combined with birth data create a rich source of data for perinatal, cardiovascular, and transgenerational health research.

Author Contributions EH, conceived the manuscript and the study and analyzed the data. MJ, conducted quality control of the linked data and the later Louisiana linkage. TS, created and merged databases from BHS and studies. DB, coordinated linkages across the three states. MW, performed initial Louisiana linkage. All authors contributed to the writing and read and approved the final manuscript.

Compliance with Ethical Standards

References

Adams, M. M., Berg, C. J., McDermott, J. M., Gaudino, J. A., Casto, D. L., Wilson, H. G., & McCarthy, B. J. (1997a). Evaluation of reproductive histories constructed by linking vital records. *Paediatric and Perinatal Epidemiology, 11*(1), 78–92. Retrieved from http://onlinelibrary.wiley.com/store/10.1111/j.1365-3016.1997.tb00799.x/asset/j.1365-3016.1997.tb00799.x.pdf?v=1&t=ipfnemup&s=79cb7876d4b7935ef8c89078af98cbea01d70621.

Adams, M. M., Wilson, H. G., Casto, D. L., Berg, C. J., McDermott, J. M., Gaudino, J. A., & McCarthy, B. J. (1997b). Constructing reproductive histories by linking vital records. *American Journal of Epidemiology, 145*(4), 339–348. Retrieved from http://aje.oxfordjournals.org/content/145/4/339.full.pdf.

Bell, R. M., Keesey, J., & Richards, T. (1994). The urge to merge: Linking vital statistics records and Medicaid claims. *Medical Care, 32*(10), 1004–1018.

Berenson, G. S. (2001). Bogalusa Heart Study: A long-term community study of a rural biracial (Black/White) population. *American Journal of Medical Sciences, 322*(5), 293–300. Retrieved from http://graphics.tx.ovid.com/ovftpdfs/FPDDNCMCGGMHDG00/fs033/ovft/live/gv017/00000441/00000441-200111000-00007.pdf.

Bonamy, A. K., Parikh, N. I., Cnattingius, S., Ludvigsson, J. F., & Ingelsson, E. (2011). Birth characteristics and subsequent risks of maternal cardiovascular disease: Effects of gestational age and fetal growth. *Circulation, 124*(25), 2839–2846. https://doi.org/10.1161/circulationaha.111.034884.

Catov, J. M., Dodge, R., Yamal, J. M., Roberts, J. M., Piller, L. B., & Ness, R. B. (2011). Prior preterm or small-for-gestational-age birth related to maternal metabolic syndrome. *Obstetrics and Gynecology, 117*(2 Pt 1), 225–232. https://doi.org/10.1097/AOG.0b013e3182075626.

Cnattingius, S., Torrang, A., Ekbom, A., Granath, F., Petersson, G., & Lambe, M. (2005). Pregnancy characteristics and maternal risk of breast cancer. *JAMA, 294*(19), 2474–2480. https://doi.org/10.1001/jama.294.19.2474.

Division of Cancer Prevention and Control. (2015). Registry Plus Link Plus. Retrieved from http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm.

Emanuel, I., Filakti, H., Alberman, E., & Evans, S. J. (1992). Intergenerational studies of human birthweight from the 1958 birth cohort. 1. Evidence for a multigenerational effect. *The British Journal of Obstetrics and Gynaecology, 99*(1), 67–74.

Freedman, M. A., Gay, G. A., Brockert, J. E., Potrzebowski, P. W., & Rothwell, C. J. (1988). The 1989 revisions of the US Standard Certificates of Live Birth and Death and the US Standard Report of Fetal Death. *American Journal of Public Health, 78*(2), 168–172. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1349110/pdf/amjph00241-0058.pdf.

Herrchen, B., Gould, J. B., & Nesbitt, T. S. (1997). Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. *Computers and Biomedical Research, 30*(4), 290–305.

Jacobs, M. B., Bazzano, L. A., Pridjian, G., & Harville, E. W. (2016). Childhood adiposity and fertility difficulties: The Bogalusa Heart Study. *Pediatric Obesity*. https://doi.org/10.1111/ijpo.12168.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*(5–7), 491–498.

Martin, J. A., & Hoyert, D. L. (2002). The national fetal death file. *Seminars in Perinatology, 26*(1), 3–11.

Nilsen, T. I., Romundstad, P. R., Troisi, R., Potischman, N., & Vatten, L. J. (2005). Birth size and colorectal cancer risk: A prospective population based study. *Gut, 54*(12), 1728–1732. https://doi.org/10.1136/gut.2004.060475.

Nitsch, D., Morton, S., DeStavola, B. L., Clark, H., & Leon, D. A. (2006). How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen Children of the 1950s study. *BMC Medical Research Methodology, 6*, 15. https://doi.org/10.1186/1471-2288-6-15.

Romitti, P. A., Watanabe-Galloway, S., Budelier, W. T., Lynch, C. F., Puzhankara, S., Wong-Gibbons, D.,.. . Alavanja, M. C. (2010). Identification of Iowa live births in the Agricultural Health Study.

Archives of Environmental and Occupational Health, 65(3), 154–162. https://doi.org/10.1080/19338241003730903.

Social Security Administration. Social security number allocations. Retrieved from https://www.ssa.gov/employer/stateweb.htm.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., & Reitsma, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology, 64*(5), 565–572. https://doi.org/10.1016/j.jclinepi.2010.05.008.

Vinikoor, L. C., Messer, L. C., Laraia, B. A., & Kaufman, J. S. (2010a). Reliability of variables on the North Carolina birth certificate: A comparison with directly queried values from a cohort study. *Paediatric and Perinatal Epidemiology, 24*(1), 102–112. https://doi.org/10.1111/j.1365-3016.2009.01087.x.