HHS Public Access

Author manuscript

J Perinatol. Author manuscript; available in PMC 2017 November 11.

Published in final edited form as:

J Perinatol. 2017 August ; 37(8): 969–974. doi:10.1038/jp.2017.70.

Evaluation of Identifier Field Agreement in Linked Neonatal Records

Eric S. Hall, PhD[1,2], **Keith Marsolo, PhD**[2], and **James M. Greenberg, MD**[1]

[1]Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[2]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

Abstract

Objective—To better address barriers arising from missing and unreliable identifiers in neonatal medical records, we evaluated agreement and discordance among traditional and non-traditional linkage fields within a linked neonatal data set.

Study Design—The retrospective, descriptive analysis represents infants born from 2013–2015. We linked children's hospital neonatal physician billing records to newborn medical records originating from an academic delivery hospital and evaluated rates of agreement, discordance, and missingness for a set of 12 identifier field pairs used in the linkage algorithm.

Result—We linked 7,293 of 7,404 physician billing records (98.5%), all of which were deemed valid upon manual review. Linked records contained a mean of 9.1 matching and 1.6 non-matching identifier pairs. Only 4.8% had complete agreement among all 12 identifier pairs.

Conclusion—Our approach to selection of linkage variables and data formatting preparatory to linkage have generalizability which may inform future neonatal and perinatal record linkage efforts.

Introduction

Despite advancements in health information exchange standards and technologies, linking of fragmented electronic health record (EHR) data representing the same individual across care settings or institutional boundaries remains challenging[1, 2, 3]. In the domain of maternal child health, these challenges are often exacerbated by the requirement for linking records representing distinct individuals (mother and child) as measures of maternal health and circumstances have direct implications on perinatal outcomes[4, 5, 6]. Further, integration of data spanning administratively distinct systems is vital to the conduct of population-based research and to the implementation of population-based data repositories[7]. While two distinct individuals represented within a data set may share a name, sex, date of birth, or

Corresponding Author: Eric Hall, 3333 Burnet Ave, ML7009, Cincinnati, OH 45229, Phone: 513-803-2083, Fax: 513-803-0968, eric.hall@cchmc.org.

Conflict of interest

The authors have no financial disclosure or conflict of interest to report.

other identifying information, the use of a sufficient set of identifiers will enable record linkage across data sets at an acceptable level of accuracy[1]. Record linkage among perinatal and neonatal records can be particularly onerous due to the absence of many identifiers traditionally used by record matching algorithms[8]. Neonatal records are typically generated at the time of, or just prior to, infant delivery, at which point identifiers such as Social Security, Medicaid, or health insurance numbers have not yet been issued[9]. In many cases, infants have not even been named. As a consequence, newborn names are often represented using the mother's surname along with temporary first names such as "Infant Girl," "Babyboy," or simply "Girl"[10, 11]. If an infant is readmitted or transferred to another care setting such as a children's hospital, yet another record is generated with identifiers that may be inconsistent with those documented in the delivery hospital setting. For example, an unnamed infant using a maternal surname at the delivery hospital may be admitted to a children's hospital with a newly given first name and the paternal surname.

Members of the current study team have previously utilized probabilistic and deterministic approaches to link perinatal data sets and evaluate specific study hypotheses[12, 13, 14, 15, 16, 17]. The present study describes initial efforts supporting the implementation of an ongoing, population-based, perinatal data repository which will facilitate evaluations spanning institutions. For example, researchers could investigate associations between measures obtained during prenatal or labor and delivery encounters with pediatric developmental outcomes measured years later at the children's hospital. During the pilot phase of the Maternal and Infant Data Hub project, regional physician billing records captured by Cincinnati Children's Hospital Medical Center (CCHMC) were integrated with neonatal records generated at a single delivery hospital, the University of Cincinnati Medical Center (UCMC). The objectives of this report are to 1) describe an approach for linking records specific to newborn populations, 2) report rates of agreement between identifier field pairs within the linked neonatal records, and 3) identify sources of discordance between those identifier field pairs. Our findings provide foundational information for developing improved strategies to link neonatal and perinatal data sets.

Methods

Study Setting

In the greater Cincinnati region, all delivery hospital nursery care is directed by a single group of physicians employed by CCHMC. Patient encounters with CCHMC neonatologists and pediatricians occurring in a delivery hospital setting generate physician billing records stored within the CCHMC EHR system. Across the greater Cincinnati region, approximately 80–90% of newborns, including all Medicaid-insured or clinically-complicated patients, receive newborn care from CCHMC physicians, generating approximately 23,000 billing records annually. Clinical coverage includes both the normal newborn nursery and the neonatal intensive care unit at UCMC, an academic medical center delivering approximately 2,500 infants each year. A separate medical record is generated within the UCMC EHR for each of these newborns. This retrospective descriptive analysis examines the cohort of infants born at UCMC during the three year period from January 1, 2013 through December

31, 2015. The CCHMC Institutional Review Board (IRB) approved this study with a waiver of informed consent. Reliance was granted by the UCMC IRB.

Data Sets

CCHMC and UCMC use versions of the Epic EHR (Epic Systems Corporation, Verona, WI) installed in 2010 and 2012 respectively; study data were retrieved from each Epic data warehouse using query logic to select infants born from 2013 through 2015. During the study timeframe, 7,792 newborn medical records were generated at UCMC, while 7,404 billing records were generated following a CCHMC physician encounter at the UCMC location. Thus, matches for all CCHMC physician billing records were expected to be found among the 7,792 newborn medical records at UCMC. However, it was not expected that all UCMC newborn medical records would have a corresponding CCHMC physician billing record for three primary reasons. First, each year approximately 100 infants were transferred from the UCMC delivery hospital to CCHMC prior to the initial physician billing charge. For these UCMC born infants, no corresponding physician billing record was ever generated at the UCMC location. Second, extremely preterm or otherwise high acuity newborns who die in the UCMC delivery room may never receive neonatal care or have a corresponding physician billing record generated (approximately 20 infants annually). Third, healthy, privately insured infants who received newborn care exclusively from non-CCHMC physicians would not have a corresponding CCHMC physician billing record; however, we are unaware of any such cases at the UCMC location during the study timeframe.

Nine variables were selected from both physician billing and newborn medical record sets for use in matching, including infant date of birth, sex, first and surname, street address and zip code, as well as birth weight, in grams. The mother's listed first and surname were also obtained from all records. In addition, paternal surname, which is captured in the CCHMC records as emergency contact information, was obtained from the physician billing records.

To process the data, all non-alphanumeric characters such as apostrophes and spaces were removed from first and surname fields. This helps in the comparison of name values. Additional data preparation methods included extraction of street number and the first word of the street name components of the street address. For example, the address "123 Phony St." would be mapped to variables "123" and "Phony." Soundex codes were also generated for infant, maternal, paternal, and street names. Soundex enables the comparison of differently spelled words by encoding words that have variations in spelling, but similar pronunciation, with the same code[18]. Finally, as birth weight may be documented at various levels of precision, we created a variable representing birth weight rounded to the nearest 10 grams.

Linkage Process

Individual records were linked using a three-step process developed previously that combines both deterministic and probabilistic components and deems records to be linked when a likelihood score threshold is exceeded[12]. The first round of linkage used the raw identifier fields found in both data sets (infant date of birth, sex, first and surname, street address, zip code, exact birth weight in grams and maternal first and surnames). All linked

records were removed prior to a second iteration in which processed identifiers replaced their raw valued counterparts (Soundex-encoded infant and maternal names, street number and Soundex-encoded street name, birth weight rounded to the nearest 10 grams). The third iteration linked records based on the similarity of delivery hospital infant and children's hospital parental (maternal and paternal) surnames (see Figure 1). After each iteration, all linked records were manually reviewed to assess the accuracy of matches.

Analysis

Within the linked data set, we calculated the number of unique values and a corresponding selectivity score for each variable. Selectivity is defined as the number of unique values divided by the total number of records, and is used to represent the variation of values for a given field. Next, we calculated the rate of agreement between neonatal records for each pair of identifier fields, as well as the missingness of values for each variable. Non-matching, or discordant fields, were defined as those in which comparison field pair values disagreed and in which neither of the comparison fields contained a missing value. Child first names which included the words "infant," "baby," "girl," or "boy," were counted as missing values and were not counted as discordant. Both raw and processed variables were listed in the agreement calculations. In our next analysis, we evaluated a final set of 12 variables which included processed rather than raw-valued variables where applicable (infant date of birth, sex, Soundex-encoded first name, Soundex-encoded surname, street number, Soundex-encoded street name, zip code, birth weight rounded to the nearest 10 grams, Soundex-encoded maternal first name, Soundex-encoded maternal surname, and comparisons of the Soundex-encoded infant surname to the Soundex-encoded maternal surname as well as the Soundex-encoded infant surname to the Soundex-encoded paternal surname), and counted the number of the 12 comparisons that were in agreement as well as the number that were discordant, not including comparisons involving missing values or temporary infant names. Finally, for each pair of raw identifier fields, we manually reviewed non-matching cases in which neither variable contained a missing value. From the review, we developed a qualitative description of frequent causes for discordance between linked records. All descriptive analyses and calculations were conducted using SAS version 9.4 (SAS Institute Inc., Cary, NC) software.

Results

As diagramed in Figure 1, from the initial set of 7,404 CCHMC physician billing records the data linkage process produced a neonatal data set including 7,293 linked record pairs (98.5%). In round one, 4,551 record pairs (61.5%) were linked using only raw identifier fields. In round two, an additional 2,111 record pairs (28.5%) were linked using processed identifier fields. Comparison of infant to parental surnames in round three resulted in an additional 631 (8.5%) linked pairs. Of the physician billing data set, 111 records (1.5%) remained unmatched. Counts of distinct values for each identifier field are listed in Table 1. Within the physician billing data set, the number of unique values ranged from two (for infant sex) to 6,976 distinct street addresses.

Table 2 presents the number of matching, non-matching, and missing values comparing corresponding identifier fields between linked records. The greatest level of agreement occurred between values for infant sex and date of birth (99.8% matching). The lowest rate of matching occurred when comparing the newborn medical record infant surname to the paternal surname in the physician billing records (29.6%). This is partially a consequence of the high rate of missing paternal surname values (48.0%). The identifier with the greatest rate of missingness was infant first name, absent from either the physician billing record, the newborn medical record, or both sources in 59.0% of the linked records.

While only 4.8% of the linked records had complete agreement among all 12 identifier pairs in the evaluation set, nearly two-thirds (65.5%) of the records contained 9 or more matching identifier pairs and 98.9% contained 5 or more matching pairs (see Table 3). More than 1 in 4 records (27.3%) contained no discordant pairs and 54.1% contained one or fewer discordant pairs. Only 4.7% of the linked records contained 5 or more discordant pairs. On average, records contained 9.1 matching (standard deviation: 1.7) and 1.6 non-matching (standard deviation: 1.4) pairs. One record contained just 2 matching pairs (sex and date of birth). In addition to missing infant first names, this record contained 8 discordant pairs including address information, similar but discordant birth weights, and incongruent infant and parental surnames. However, although name spelling was computationally dissimilar when comparing both raw-valued and Soundex-encoded field pairs, the veracity of the links was easily established by manual review.

Results of the qualitative review of discordant identifier field pairs are summarized in Table 4. Discordance in many of the fields resulted from inconsistent spelling as well as apparent clerical entries during data entry or transcription. Another common theme was the use of aliases, or nick-names, such as “Katie” instead of “Catherine” in one of the source records. Among discordant birth weights, the median difference between values was just 1 gram (mean difference of 20.1 grams).

Discussion

Our linkage strategy identified matches for 98.5% of neonatal physician billing records. Also, our analysis of the linked records provided several insights of great potential value for informing future efforts involving perinatal populations. Within our linked study data set, only 2,602 of 7,293 records (35.7%) had complete agreement (deterministic matching) between infant date of birth, sex, first name, and surname comparing data sources. The use of Soundex encoding of infant names provides only a modest boost to 2,697 records (37.0%). A primary reason for the low match rate using these criteria only is the high rate of missingness within the child first name field. This finding reemphasizes the previously described utility of a probabilistic, in combination with a deterministic, strategy for linking perinatal records[8]. Additionally, our findings demonstrate that inclusion of non-traditional linkage fields such as birth weight and parental names may further enhance a probabilistic approach. Birth weight is particularly useful within the neonatal domain as it provides a mechanism to distinguish between twin babies who have a great deal of overlap in identifiers including date of birth, surname, parental names, and address information. In fact, when permanent first names have not yet been given, birth weight is one of the few ways to

differentiate between same sex twins, as even identical twins are unlikely to have exactly equal birth weights[19]. Birth weight is also nearly universally measured and documented at the time of birth, making it highly useful for record linking. Both record sets also captured gestational age. However the measure was represented within the physician billing set using complete weeks with values ranging from 23 to 44 – the vast majority of which contained values from 37 to 40. Due to the limited granularity and the inability to differentiate between twins, we chose not to include gestational age as a non-traditional linkage field in this evaluation. Finally, while the street address field alone had a high rate of discordance (approximately 50%), the street number and Soundex-encoded first street name fields each agreed in approximately 80% of cases greatly increasing the contribution of address information to linkage algorithms.

Many of the data preparation tactics we employed addressed problems identified in the qualitative assessment of discordant fields. Soundex encoding and the extraction of street address components helped to mitigate the effects of subtle inconsistencies in spelling. Rounding birth weights to the nearest 10 grams lessened the effect of 1 or 2 gram discrepancies in documented infant birth weights. While less than half (48.2%) of exact birth weights matched comparing linked records, more than 80% matching was achieved when birth weights were rounded to the nearest 10 grams. Other causes, such as miskeyed values or transcription errors are more difficult to address with a data preparation or linkage strategy. Finally, complete documentation of paternal information in both records would substantially aid in linkage. Documented maternal surnames have several opportunities for discordance, including the inconsistent use of maiden names or the improper replacement with the father's surname when mother has a different surname. Not only would the availability of paternal first and surname fields provide additional linkage variables, but paternal surnames are less likely to vary between data sources and may prove to be a more stable identifier than maternal surnames. Of course, obtaining complete paternal information has its own challenges and may be more complex than obtaining maternal information, particularly in a labor and delivery or newborn nursery setting. Use of more distinctive temporary infant names incorporating maternal first names (e.g. "Elizabethsgirl" versus "Girl") is another potential mechanism to improve patient identification and record linkage[10].

Although the present analysis represents efforts within a single hospital setting, there are many generalizable methods that may inform the efforts of other perinatal researchers. The described data preparation approach for improved matching between inconsistently spelled names is one useful tactic recommended for implementation in other settings. While the current study used Soundex encoding, other methods such as Metaphone, which is an improvement over the Soundex phonetic algorithm[20], or the Jaro-Winkler method which compares individual letters in name fields for similarity[21], can also be employed. Based on our linkage success rate and evaluation, however, we would expect any additional improvements to be marginal. Nevertheless, the comparison of various name processing approaches could be the subject of future research efforts. Our decision to compare parental to infant surnames is another recommended tactic for future efforts, whether they be the linkage of neonatal data sets or the linkage of other perinatal data sets, such as linkage between maternal and child records. Finally, in this study approximately 99% of records had

5 or more matching field pairs and 5 or fewer non-matches. While other researchers may choose the subset of linkage variables appropriate to their setting, identifying the thresholds at which matching and/or non-matching field pairs would trigger a manual review could aid in automation of the linkage process. For example, using our 12 linkage variables, records with fewer than 5 matching pairs or greater than 5 non-matching pairs would be flagged for manual review, while records with a greater number of matched variables or fewer non-matching variables would be exempt. Researchers could adjust these thresholds to allow for tolerance for mismatched records that is acceptable for their specific study.

The study utilized records obtained from a single care setting, which means that some of the data may reflect the data capture idiosyncrasies particular to that environment. Thus, there is potential for some variation in linkage and variable discordance rates if using records originating from another institution or geography. Nevertheless, the described approach has general relevance and applicability beyond the specific data sets used in the current evaluation. Also, as the focus of our analysis was on a validated, linked data set, we did not evaluate the characteristics of the unmatched records. In addition, no efforts were made to interpolate values for missing data elements.

Conclusions

Strategies for linking neonatal records must take into consideration the distinctive characteristics of the relevant data sets, such as the absence of traditional identifiers used for record linking exercises and the availability of differentiating measures such as birth weight. Our approach to the selection of linkage variables and to data preparation have broad generalizability to the linkage of perinatal data sets beyond the example described in this study. Our analysis of the agreement and discordance among linked records also provides critical insights, which may inform future efforts to link neonatal and perinatal data sets.

Acknowledgments

Funding Source

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health through the Center for Clinical and Translational Science and Training at the University of Cincinnati [5UL1TR001425-02] and the Cincinnati Children’s Research Foundation Academic and Research Committee. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014; 311(24):2479–2480. [PubMed: 24854141]
2. Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. Information Systems. 2013; 38(6):946–969.
3. Li B, Quan HD, Fong A, Lu MS. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Serv Res. 2006; 6
4. Dufendach KR, Lehmann CU. Topics in Neonatal Informatics: Essential Functionalities of the Neonatal Electronic Health Record. Neoreviews. 2015; 16(12):e668–e673.

5. Delnord M, Szamotulska K, Hindori-Mohangoo AD, Blondel B, Macfarlane AJ, Dattani N, et al. Linking databases on perinatal health: a review of the literature and current practices in Europe. Eur J Public Health. 2016; 26(3):422–430. [PubMed: 26891058]
6. 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. Paediatric and Perinatal Epidemiology. 1997; 11(S1):5–22. [PubMed: 9018711]
7. Kotelchuck M, Hoang L, Stern JE, Diop H, Belanoff C, Declercq E. The MOSART database: linking the SART CORS clinical database to the population-based Massachusetts PELL reproductive public health data system. Matern Child Health J. 2014; 18(9):2167–2178. [PubMed: 24623195]
8. Baldwin E, Johnson K, Berthoud H, Dublin S. Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. Pharmacoepidemiol Drug Saf. 2015; 24(1):45–51. [PubMed: 25408418]
9. Spooner SA, Council on Clinical Information Technology AAoP. Special requirements of electronic health record systems in pediatrics. Pediatrics. 2007; 119(3):631–637. [PubMed: 17332220]
10. Adelman J, Aschner J, Schechter C, Angert R, Weiss J, Rai A, et al. Use of Temporary Names for Newborns and Associated Risks. Pediatrics. 2015; 136(2):327–333. [PubMed: 26169429]
11. Gray JE, Suresh G, Ursprung R, Edwards WH, Nickerson J, Shiono PH, et al. Patient misidentification in the neonatal intensive care unit: quantification of risk. Pediatrics. 2006; 117(1):e43–47. [PubMed: 16396847]
12. Hall ES, Goyal NK, Ammerman RT, Miller MM, Jones DE, Short JA, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Matern Child Health J. 2014; 18(1):316–325. [PubMed: 23420307]
13. Seske LM, Muglia LJ, Hall ES, Bove KE, Greenberg JM. Infant Mortality, Cause of Death, and Vital Records Reporting in Ohio, United States. Matern Child Health J. 2016:1–7.
14. Hall ES, Venkatesh M, Greenberg JM. A population study of first and subsequent pregnancy smoking behaviors in Ohio. J Perinatol. 2016; 36(11):948–953. [PubMed: 27467563]
15. Goyal NK, Folger AT, Hall ES, Ammerman RT, Van Ginkel JB, Pickler RS. Effects of home visiting and maternal mental health on use of the emergency department among late preterm infants. J Obstet Gynecol Neonatal Nurs. 2015; 44(1):135–144.
16. Goyal NK, Hall ES, Jones DE, Meinzen-Derr JK, Short JA, Ammerman RT, et al. Association of maternal and community factors with enrollment in home visiting among at-risk, first-time mothers. Am J Public Health. 2014; 104(Suppl 1):S144–151. [PubMed: 24354835]
17. Goyal NK, Hall ES, Meinzen-Derr JK, Kahn RS, Short JA, Van Ginkel JB, et al. Dosage effect of prenatal home visiting on pregnancy outcomes in at-risk, first-time mothers. Pediatrics. 2013; 132(Suppl 2):S118–125. [PubMed: 24187113]
18. Russell R, Odell M. Soundex. US Patent. 1918; 1
19. Cheung VY, Bocking AD, Dasilva OP. Preterm discordant twins: what birth weight difference is significant? Am J Obstet Gynecol. 1995; 172(3):955–959. [PubMed: 7892890]
20. Philips L. Hanging on the metaphone. Computer Language. 1990; 7(12)
21. Winkler, WE. The state of record linkage and current research problems. US Census Bureau; 1999. Statistical Research DivisionCiteseer; 1999

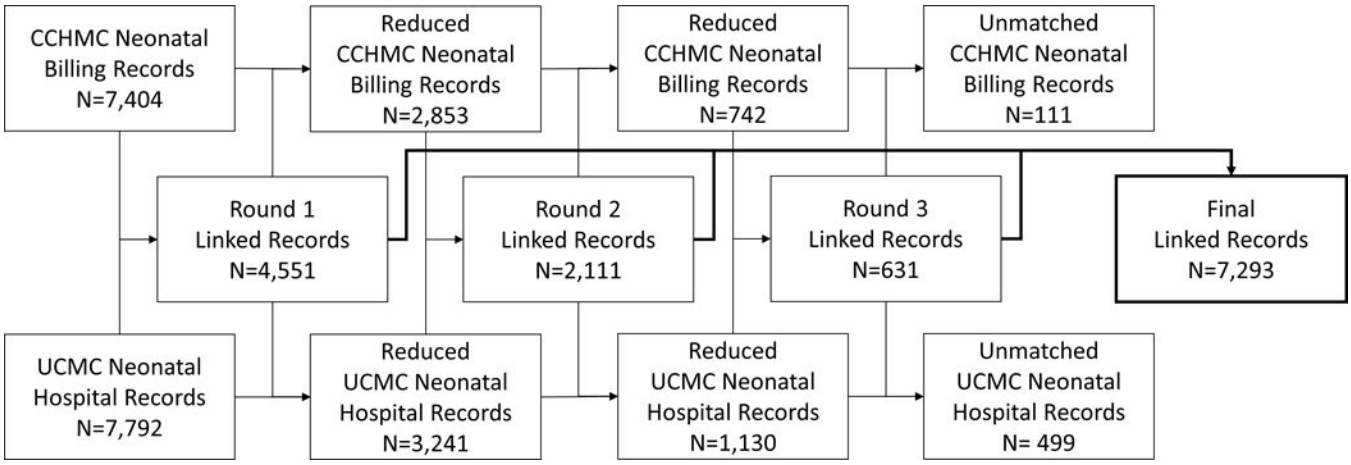

Figure 1.
Data linkage flow diagram.

Table 1

Count of distinct values and selectivity of each identifier field within the set of 7,293 linked records.

Identifier Field	Distinct Values in the Physician Billing Record Set	Selectivity in the Physician Billing Record Set	Distinct Values in the Newborn Medical Record Set	Selectivity in the Newborn Medical Record Set
Infant Sex	2	0.0%	2	0.0%
Zip Code	270	3.7%	277	3.8%
Birth Weight (Nearest 10 Grams)	430	5.9%	429	5.9%
Mother First Name (Soundex-Encoded)	1,022	14.0%	902	12.4%
Date of Birth	1,092	15.0%	1092	15.0%
Infant First Name (Soundex-Encoded)	1,103	15.1%	889	12.2%
Father Surname (Soundex-Encoded) *	1,494	20.5%	0	0.0%
Street Name (Soundex-Encoded)	1,765	24.2%	1,730	23.7%
Birth Weight (Exact)	2,089	28.6%	1,709	23.4%
Mother Surname (Soundex-Encoded)	2,367	32.5%	2,408	33.0%
Father Surname *	2,380	32.6%	0	0.0%
Infant Surname (Soundex-Encoded)	2,429	33.3%	2,471	33.9%
Street Name	2,653	36.4%	2,533	34.7%
Mother First Name	3,120	42.8%	2,808	38.5%
Infant First Name	3,164	43.4%	2,305	31.6%
Mother Surname	3,828	52.5%	3,834	52.6%
Street Number	3,950	54.1%	3,912	53.6%
Infant Surname	3,956	54.2%	3,977	54.5%
Street Address	6,976	95.7%	6,687	91.7%

* Father's surname was not available in the newborn medical record.

Table 2

Number of matching, non-matching, and missing values for each identifier field within the set of 7,293 linked neonatal records.

Identifier Field	Matching N, %	Non-Matching N, %	Missing from Only Physician Billing Records N, %	Missing from Only Newborn Medical Records N, %	Missing from Both Physician Billing and Newborn Medical Records N, %
Infant Sex	7,281, 99.8%	11, 0.2%	0, 0.0%	1, 0.0%	0, 0.0%
Date of Birth	7,280, 99.8%	13, 0.2%	0, 0.0%	0, 0.0%	0, 0.0%
Mother Surname (Soundex-Encoded)	6,945, 95.2%	270, 3.7%	72, 1.0%	5, 0.1%	1, 0.0%
Mother Surname	6,865, 94.1%	350, 4.8%	72, 1.0%	5, 0.1%	1, 0.0%
Mother First Name (Soundex-Encoded)	6,376, 87.4%	100, 1.4%	71, 1.0%	745, 10.2%	1, 0.0%
Mother First Name	6,276, 86.1%	200, 2.7%	71, 1.0%	745, 10.2%	1, 0.0%
Zip Code	6,265, 85.9%	867, 11.9%	0, 0.0%	161, 2.2%	0, 0.0%
Street Number	6,023, 82.6%	1,109, 15.2%	0, 0.0%	161, 2.2%	0, 0.0%
Birth Weight (Nearest 10 Grams)	5,963, 81.8%	971, 13.3%	228, 3.1%	120, 1.6%	11, 0.2%
Street Name (Soundex-Encoded)	5,793, 79.4%	1,300, 17.8%	16, 0.2%	182, 2.5%	2, 0.0%
Street Name	5,642, 77.4%	1,451, 19.9%	16, 0.2%	182, 2.5%	2, 0.0%
Mother or Father Surname (Physician Billing) to Infant Surname (Newborn Medical) (Soundex-Encoded)	5,211, 71.5%	2,020, 27.7%	62, 0.9%	0, 0.0%	0, 0.0%
Mother or Father Surname (Physician Billing) to Infant Surname (Newborn Medical)	5,146, 70.6%	2,085, 28.6%	62, 0.9%	0, 0.0%	0, 0.0%
Infant Surname (Soundex-Encoded)	5,047, 69.2%	2,246, 30.8%	0, 0.0%	0, 0.0%	0, 0.0%
Infant Surname	5,003, 68.6%	2,290, 31.4%	0, 0.0%	0, 0.0%	0, 0.0%
Mother Surname to Infant Surname (Soundex-Encoded)	4,212, 57.7%	3,008, 41.2%	73,1.0%	0, 0.0%	0, 0.0%
Mother Surname to Infant Surname	4,173, 57.2%	3,047, 41.8%	73, 1.0%	0, 0.0%	0, 0.0%
Birth Weight (Exact)	3,516, 48.2%	3,418, 46.9%	228, 3.1%	120, 1.6%	11, 0.2%
Street Address	3,437, 47.1%	3,695, 50.7%	0, 0.0%	161, 2.2%	0, 0.0%
Infant First Name (Soundex-Encoded)	2,898, 39.7%	91, 1.2%	806, 11.1%	2,700, 37.0%	798, 10.9%
Infant First Name	2,803, 38.4%	186, 2.6%	806, 11.1%	2,700, 37.0%	798, 10.9%
Father Surname (Physician Billing) to Infant Surname (Newborn Medical) (Soundex-Encoded)	2,210, 30.3%	1,585, 21.7%	3,498, 48.0%	0, 0.0%	0, 0.0%
Father Surname (Physician Billing) to Infant Surname (Newborn Medical)	2,159, 29.6%	1,636, 22.4%	3,498, 48.0%	0, 0.0%	0, 0.0%

Table 3

The number of matching as well as non-missing, non-matching identifier pairs within the set of 7,293 linked neonatal records. *

Count of Matching Pairs	Number of Linked Records (N)	Absolute Percent (%)	Cumulative Percent (%)	Count of Non-Matching Pairs	Number of Linked Records (N)	Absolute Percent (%)	Cumulative Percent (%)
12	347	4.8%	4.8%	12	0	0.0%	0.0%
11	1,310	18.0%	22.7%	11	0	0.0%	0.0%
10	1,738	23.8%	46.6%	10	0	0.0%	0.0%
9	1,382	18.9%	65.5%	9	0	0.0%	0.0%
8	1,228	16.8%	82.3%	8	2	0.0%	0.0%
7	714	9.8%	92.1%	7	3	0.0%	0.1%
6	308	4.2%	96.4%	6	67	0.9%	1.0%
5	189	2.6%	98.9%	5	272	3.7%	4.7%
4	60	0.8%	99.8%	4	532	7.3%	12.0%
3	16	0.2%	100.0%	3	748	10.3%	22.3%
2	1	0.0%	100.0%	2	1,724	23.6%	45.9%
1	0	0.0%	100.0%	1	1,954	26.8%	72.7%
0	0	0.0%	100.0%	0	1,991	27.3%	100.0%

* The 12 identifier pairs included in the evaluation were: infant date of birth, infant sex, Soundex-encoded infant first name, Soundex-encoded infant surname, street number, Soundex-encoded street name, zip code, infant birth weight rounded to the nearest 10 grams, Soundex-encoded maternal first name, Soundex-encoded maternal surname, comparison of the Soundex-encoded infant surname to the Soundex-encoded maternal surname, and comparison of the Soundex-encoded infant surname to the Soundex-encoded paternal surname.

Table 4

Qualitative summary of causes for discordance between identifier field pairs within linked records.

Identifier Field	Qualitative Assessment
Birth Weight (Exact)	Birth weights disagreed by one or two grams Miskeyed, missing, or transposed digits in one record
Date of Birth	"Day" component of date disagreed by one or two days Transposed digits within the "Day" component
Infant First Name	Transposed first and surnames in one record Inconsistent name spelling Use of an alias in one record
Infant Surname	Inconsistent name spelling Use of hyphenated parental surname in only one record Matches maternal surname in one record and paternal in the other
Infant Sex	Incorrectly coded value in one record, in several cases a child with a first name like "Infant Boy" was assigned "Female" sex within the same record.
Mother First Name	Inconsistent name spelling Use of an alias in one record
Mother Surname	Inconsistent name spelling Use of father's rather than mother's surname in one record
Street Address	Completely different addresses listed Inconsistent spelling of street names Inconsistent street suffixes (e.g. "Boulevard" versus "Blvd." or "Street" versus "Drive") Inconsistent formatting of apartment abbreviations (e.g. "Apt 3" versus "#3") Miskeyed or transposed digits in street number of one record
Zip Code	Miskeyed or transposed digits in one record Erroneous zip code in one record (street addresses match, one zip code is incorrect)