Creation of a linked inter-agency data warehouse: the Longitudinal Study of Early Development

A Research Report from the New York City Department of Health and Mental Hygiene

Melissa Pfeiffer, MPH
Meredith Slopen, MSW
Allison Curry, PhD
Katherine McVeigh, PhD, MPH

Creation of a linked inter-agency data warehouse: the Longitudinal Study of Early Development

Abstract

Government agencies are increasingly seeking to leverage existing administrative data for research and policy-making. However, since linkage methodologies are rarely published, individual jurisdictions must each develop techniques in isolation. To facilitate future projects and consistency across sites, the authors describe methods and results of linking child health and education data, including data preparation, threshold setting, quality assurance, false match rates, comparisons to expected yields, and cohort characteristics. A probabilistic approach was used to link children and siblings across 5 sources: the New York City Department of Health and Mental Hygiene's Early Intervention Program (n=156,834), Lead Poisoning Prevention Program Registry (n=1,469,265), Birth Certificate Registry (n=1,380,608), and Death Certificate Registry (n=8,331), and the New York City Department of Education's administrative, special education and testing databases (n=617,934). The resulting relational data warehouse, the Longitudinal Study of Early Development, contains data for a diverse population of 1,942,942 children born 1999−2004, with a 0.6% estimated false match rate. Over half (57%) of the children were found in more than 1 source; 20% had at least 1 sibling. Techniques developed for this project may be replicated by other jurisdictions to link sources and answer important policy questions.

Introduction

Given financial and logistic challenges posed by conducting primary data collection, the public health research community is increasingly linking routinely collected surveillance and administrative data to address research and policy issues (1). Linkage projects, though challenging, result in data sets with valuable advantages. They have the capacity to provide an extensive, population-level, longitudinal data source appropriate for vital research inquiries and subgroup analyses that would not be possible with any 1 source (1). While strategies and parameters like acceptable levels of error will vary, all such projects must incorporate assessments of linkage results and those with large data sets will involve automation. However, to our knowledge, detailed descriptions of employed probabilistic data linkage techniques have rarely been published in the epidemiologic literature (2−5). Consequently, jurisdictions undertaking such projects have been developing procedures in isolation.

Many important questions raised in administrative settings such as public health departments, including those pertaining to the health and development of children, cannot be answered with any single available data source. In New York City, this led to collaboration among several programs within the Department of Health and Mental Hygiene (DOHMH) and the Department of Education to create the Longitudinal Study of Early Development (LSED) data warehouse. By describing the innovative techniques developed to simultaneously link 5 data sources and create the LSED data warehouse, we aim to enable other jurisdictions to leverage existing data to expand the availability of public health data and knowledge.

Materials and Methods

Our primary goal was to create a data warehouse containing linked data from 5 sources: (a) the DOHMH Birth Certificate Registry (Birth); (b) the DOHMH Early Intervention Program's administrative database (EI); (c) the DOHMH Lead Poisoning Prevention Program Registry (Lead); (d) Department of Education administrative, special education, and testing databases (Education); and (e) the DOHMH Death Certificate Registry (Death). Secondary goals were to link siblings born to the same mother and to join residential records within and across sources (methods for the latter are not described here). We describe the data sets, linkage strategies, quality assurance evaluations, and final warehouse preparation steps. We use the term *linkage* to refer to the process of bringing records together for evaluation, whereas *match* refers to the outcome of joined records (see Table 1 for definitions; terms italicized when introduced).

The plan to construct the LSED data warehouse was agreed upon through a Memorandum of Understanding between the New York City DOHMH and the Department of Education and through a data sharing agreement among DOHMH bureaus participating in the project. The DOHMH Institutional Review Board approved the use of the LSED data warehouse for specified research purposes. The project team included investigators with expertise in the source data set content and technical experts with experience in conducting linkage projects. IBM's QualityStage 8.0 probabilistic linkage software, designed to identify records belonging to a single entity within 1 data source or between multiple data sources, was used to standardize and link data.

Data preparation and structure

Identifiers were standardized before the linkage began (e.g., "Bob," "Bobbie," and "Rob" were converted to "Robert" and "St." became "Street"). Data sets differed with respect to the presence of *unresolved duplication* (more than 1 identification number for a child) and the occurrence of multiple observations per identification number, as well as in the identifiers themselves (Table 2). A critical assumption for the process was that the Birth and Death registries had no unresolved duplication, i.e., each child identification number represented a unique child. However, some children in the EI, Lead, and Education data sets had more than 1 identification number. Further, the EI and Lead data sets were structured to contain multiple records for the

same identification number in which alternate identifying information was documented. To capitalize on this variant information, we used all records in the EI and Lead data sets; consequently, the number of observations is greater than the number of unique children in those sources.

Creation of passes

We conducted the child linkage in 2 *phases* and the sibling linkage in 1 phase. Although each phase was executed with slightly different records, linkage criteria, and review methods, all phases involved an iterative process of creating multiple passes, with 2 stages for each pass. The first stage, *blocking*, creates groups of records (blocks) with identical values for selected variables or parts of variables. In the second stage, records within the same block are compared and joined based on values of additional variables, with allowances for similar but not exact values. For example, 1 pass created blocks of records with identical values on components of the child's first and last name, ZIP code, and date of birth, and then compared records within each block on child's full name, address, sex, social security number, father's name, and mother's name and date of birth. Blocks are small relative to the total number of records to be evaluated, thereby improving efficiency by substantially reducing the number of comparisons.

Assignment of weights

After the execution of each pass, records identified as potential matches were grouped into a set with a single identification number; the most complete record in the set was selected by the software as the *primary record*, and the others were marked as *secondary records*. Each secondary record was given a *weight*, which is a positive numeric value reflecting the similarity of the records brought together. Higher weights indicate greater likelihood that the secondary record represents the same child as the primary record in the set.

Review of potential matches

The first phase of the child linkage compared the 2 data sets containing unique children—Birth and Death. This was restricted to non-infant deaths; results of routine linkages with birth certificates were used for infant deaths. Due to the analytic importance of associating a child

with a death record and the relatively small number of potential match sets, every set generated in this phase was evaluated by independent reviewers.

The second phase of the child linkage included all 5 data sources combined into a single file. For the sibling linkage, sibling sets, defined as groups of children born to the same mother, were identified within the single file of all records. However, Death and Education records were not included in this process as they did not contain mother's information (Death) or contained guardian information potentially related to the father or another adult (Education). The search for matches within a single file of all records maximized the potential for discovering matches between records in which only a portion of the identifiers were the same by incorporating additional records that contained the discrepant information. Figure 1 depicts the process using hypothetical data. In Figure 1A, 5 records from different original sources are collapsed into a single data set. The first 3 records (numbers 23, 42, and 57) have identical values on the blocking criteria and are consequently evaluated as potential matches (Figure 1B). Records 23 and 42 are identified as the same person, as are records 42 and 57. Records 23 and 57 belong to the same unique child because record 42, having date of birth in common with record 23 and address in common with record 57, bridged the information. Since none of the data sets were used as the source to which all other records were compared, matches were created between all combinations of sources. Specifically, since we did not simply link each source against Births, we were able to identify children with records in Lead and Education who were not born in New York City.

Due to the volume of potential matches generated, only samples of potential matches produced in the second phase of the child linkage and in the sibling linkage were examined. For each pass, we sampled approximately 800 secondary records that replicated the weight distribution of all secondary records in the pass. The sample was then reviewed by 2 independent reviewers with tie decisions broken by a third reviewer. Although only the sampled secondary record was scored to indicate whether it truly represented the same child as the primary record, all secondary records within the set were visible, allowing reviewers to consider various permutations of identifiers that may exist for a single child. *Reviewer concordance*—the percentage of records on which reviewers agreed—was assessed for each sample. Low initial concordance on an exploratory sample prompted a discussion of the relative importance of identical or differing

information and the minimum amount of identical information necessary for a match; general guidelines for decisions were then formed. For example, one guideline indicated that 2 records with identical first name, last name, and date of birth could be deemed 1 person by the reviewers, and another indicated that although identical addresses could indicate 1 person, differing addresses did not indicate different people, given that residents are known to move within the city. High concordance (>80%) indicated that subjective human reviews were using similar and consistent criteria for decisions.

Setting threshold weights

Information from these samples was then used to establish *threshold weights* for each pass. The threshold is the weight above which all secondary records are automatically accepted and below which they are rejected as *non-matches*, or records belonging to more than 1 person. Once sampled secondary records were scored, *cumulative false match rates* were calculated for each weight as the proportion of *false matches* (records incorrectly joined as 1 person) among all sampled secondary records with the same or higher weight. Our threshold was the lowest weight at which the cumulative false match rate was below 1%. We selected 1% as the proportion of false to allow for maximal linkages while limiting the number of inappropriately joined individuals.

Finalizing linkages

Upon completion of both phases of the child linkage, the accepted matches from each pass were combined to establish unique children and their associated records. Similar to the single file approach allowing 2 records to be linked within a pass because of a common third record, combining accepted matches from all passes allowed 2 records that were not joined within 1 pass to be identified as those of a single child because both records were joined with the same third record, albeit in different passes. In Figure 1B, pass 1 identified records 23, 42, and 57 as belonging to a single child. In pass 2 (Figure 1C), which used different criteria, records 42 and 68 were linked. When results of all passes were combined (Figure 1D), records 23, 42, 57, and 68 were identified as those of a single child because they were all linked, in different passes, with record 42. As a result of each record participating in each pass, it was possible at this stage for a set of records determined to belong to a single unique child to have more than 1 birth or

death record. Any set that contained more than 1 birth or death record was reviewed by a project investigator, who separated the set of records into distinct children. For instance, if a set contained more than 1 birth record, the investigator determined whether each of the non-birth records truly matched 1 of the birth records, neither birth record but 1 of the other records, or none of the records within the set.

Evaluating linkages

Once sets of records for unique children were finalized, we extracted a sample to calculate the overall false match rate using 2 denominators: 1) all unique children, and 2) the subgroup of unique children who had more than 1 original identification number linked together through this process, (i.e., at least 2 different records represented a unique child). As with extractions for threshold setting, we sampled approximately 800 unique children for the subgroup. To maintain proportions similar to the entire cohort, the sample of all children contained nearly 1,400 children. We verified that the data sources were proportionally represented in the sample. For each sampled child with more than 1 original identification number, reviewers examined all identifiers and determined whether they corresponded to 1 child. We also determined the source-specific false match rate by determining the proportion of false matches among sampled children with at least 1 record from the source.

We followed a similar process to estimate the final sibling false match rate. As further evaluation, we calculated the proportion of sibling sets that had an *invalid birth interval*, defined as a sibling set that had at least 2 children with birth dates more than 2 days but less than 7 months apart. This interval allowed multiple births to be born on different days, such as before and after midnight, as well as premature delivery of a subsequent pregnancy.

Final preparation

Once the final determination of unique children was made, we resolved discrepant information across data sources for analytic variables that could have only 1 value. These variables included month and year of birth for the child and mother, and the sex of the child. For each of these variables, we prioritized the values from Birth, the source we considered most accurate. If there was no birth record, the next most trusted source was Death. If there were no data from that

source, we used the most frequent value among the remaining sources. Rather than selecting 1 value based on source or prevalence, race/ethnicity for the child was combined from all 4 sources in which it was recorded (Lead, EI, Education, and Death) to create 6 variables, each of which represents 1 race/ethnicity category and has values of yes or null/missing. In this report, the 6 variables are collapsed into 1 according to the following hierarchy: Hispanic, if Hispanic in any source, regardless of other values; multiple race, if more than 1 non-Hispanic variable (i.e., black, white, Asian/Pacific Islander, or other) is yes; black, white, Asian/Pacific Islander or other if only that variable is yes; unknown if that variable is yes (i.e., no race/ethnicity information was found in any source).

Results

Child linkage outcomes

We linked 1,806 non-infant death records against the 1,380,608 birth records during phase 1 of the child linkage. There were 1,357 secondary records identified from all passes; 1,299 (95.7%) were accepted by the reviewers as matches, with reviewer concordance averaging 98.6%.

In 5 of the 9 passes created for phase 2 of the child linkage, all potential matches were accepted because the cumulative false match rate never rose above 1%. In the remaining 4 passes, threshold weights were established. Reviewer concordance was again very high, varying from 82.6% to 100.0%. After combining records from all accepted matches, there were 30 children who had more than 1 death record and 384 who had more than 1 birth record. After final resolution of these sets, the combined data sets encompassed records for 1,942,942 children.

Cohort characteristics

The proportion of male and female children was nearly equal (Table 3). More than half of children (57.4%) did not have race/ethnicity information from any of the 4 sources that report child race. Most children with race/ethnicity data were Hispanic, followed by black (non-Hispanic) and white (non-Hispanic). The distribution of mother's race/ethnicity, available from the birth certificate, was slightly different, with white (non-Hispanic) being the second most common group rather than black (non-Hispanic), and less than one-third (29.2%) unknown. When the child's race/ethnicity and mother's race/ethnicity were both known, they were the same 87.6% of the time.

The number of children born decreased slightly each year: 11.2% of the population was born in 1994, and 7.6% in 2004. The majority of children (73.0%) had a lead record; nearly the same proportion (71.1%) had a birth record. More than 1 million had records from multiple sources. Among those with a record in more than 1 source, the most common combinations were Birth and Lead (44.4%) followed by Birth, Lead, and Education (29.2%) (Table 4).

Evaluations

The overall false match rate among children was 0.4%. The source-specific rates ranged from 0.0% (Death) to 0.9% (Education), although these figures likely represent the maximum false match rate for each source, since the falsely matching records might not be from that source. After removing from the denominator children who had only 1 original identification number and therefore could not be false matches, the overall false match rate was 0.6%, and source-specific rates ranged from 0.0% (Death) to 1.1% (Education).

The matrix of matches between our data sources can be found in Table 5, which shows, for each source data set, the proportion of unique children in that data source (rows) with a match in the other data sources (columns). Within the EI, Lead, and Education data sets, 0.6%, 3.4%, and 1.3% of the children, respectively, linked with another record within that data set. These proportions represent duplication within these sources, or the proportion of unique children who had more than 1 identification number within the same data set.

To assess the appropriateness of thresholds established to distinguish accepted matches from non-matches, we evaluated 10 of the Table 5 cells against our expectations of the yield from this process. These expectations were derived from previous linkages conducted for different purposes with different software as well as existing information obtained from external sources including the proportion of foreign-born children in the Lead registry and in public schools, which indicate the minimum proportion of children in Lead and Education who would not be expected to have a record in Birth. The yield in 7 of the cells (children in Birth, EI, Lead, and Education linked with Birth; children in EI with matches in Lead; and children in EI and Education linked with Education) was within the range of our expectations. Two of the cells (duplication within Lead and within EI) were below our expected match rate, and 1 cell (children in Education with a match in Lead) was above.

Sibling characteristics and evaluations

Most of the children (63.2%) do not have a sibling in the data warehouse; however, the sibling linkage did identify 306,240 sets of children born to the same mother. The majority of sibling sets (75.1%) had 2 children, with a maximum of 14. Of the 800 final sibling sets sampled, 13

(1.6%) were determined to be false matches. Less than 1% of sibling sets had an invalid birth interval, consistent with our sibling false match rate.

Discussion

We created the Longitudinal Study of Early Development (LSED) data warehouse to maximize the value of existing data sources and investigate critical questions related to early child development. The completed LSED data warehouse contains records from 5 data sources for a diverse population of 1,942,942 unique children and 306,240 sibling sets, with false match rates of <1% and 1.6%, respectively.

The construction of the LSED data warehouse required the development of novel techniques, including sampling secondary records and threshold setting, implemented by staff who were experienced with data linkages and had time allocated for this project. Other jurisdictions may need to invest in software and staff resources and should expect to experience steep learning curves with the first endeavor. Our aim is to share our methods, which can be applied to other populations with adjustments for different purposes and data. For example, the purpose influences the tolerance of false matches. Research projects may tolerate more false matches than surveillance projects, if data for the latter are used for patient follow-up or linkage is on-going, since it is easier to connect two records to 1 person than to separate records from 2 incorrectly joined individuals upon receipt of additional information. Many projects involve the merger of 2 data files, but linkages of more than 2 data files require choosing between a simultaneous linkage of all sources and a series of sequential linkages. We believe that the ability to leverage incomplete connections inherent within a simultaneous process allowed for the discovery of additional matches. Additionally, fewer iterations of the process are necessary when all sources are linked simultaneously.

Recommendations for others based on the strengths and limitations of our strategies fall into 3 main categories: data concerns, process or efficiency issues, and aspects of evaluation. Our first data recommendation is that those executing the linkage should be well-informed about variable accuracy and meaning. In our project, although some variables were expected to be accurate, like date of birth in the Birth Certificate Registry, other variables contained unexpected or less certain information. For instance, in some cases the value for child's social security number may have in fact belonged to a parent. Guardian name in Education could have been the name of the mother, father, or another guardian (e.g. a grandparent). It is important that staff responsible for creating

the linkage understand data elements available; this project benefited by having staff experienced with source data involved with linkage development. Second, allowing multiple observations per original child identification number from EI and Lead as we did may have complicated execution and generated extraneous secondary records. We recommend searching for matches on multiple name and address fields, such as comparing address in 1 record with either primary address or secondary address in another record, rather than either searching different records or selecting one as the best candidate.

Projects we undertake in the future will benefit from a more streamlined process. First, the expectations for the yield of matches our linkage would produce were agreed upon later in the process, but they should have been formed before the linkage began based on existing information about the population and prior linkage projects. This would have permitted monitoring progress as passes were completed, providing us with guidance as to whether the yield was too low due to conservative match criteria or too high, indicating that a pass should have been revised with more restrictive criteria. Additionally, establishing these expectations earlier would have assured that they were not influenced by the yield achieved. However, we believe that 9 out of 10 cells fell within or below expectations; therefore, we were more likely to be under-matching than over-matching. Second, it may have been possible to achieve the same results without executing the Birth-Death linkage as a separate phase. Given the importance and finality of a date of death, in the future we would examine each potential match set that contains a death record individually. The number of such sets would likely be manageable.

Evaluations are necessary whenever decisions are automated. Our formal measures of evaluation were limited to false match rates and invalid birth interval (applicable only to sibling linkages). The acceptable false match rate should be agreed upon, ideally before beginning the linkage, based on the purpose of the linked data and the tolerance for both false matches and *false non-matches* (i.e., records that were not matched but in fact belong to the same person). We were willing to accept a small proportion of false matches (up to 1% for children) based on the intended use of the data warehouse, which is research rather than surveillance. Allowing some false matches reduced the number of false non-matches. We did not determine the proportion of false non-matches; as a proxy, we calculated the proportion of true matches among all sampled secondary records below the threshold weights and found the proportions to be acceptably small.

A system for estimating a false non-match rate is currently being developed by the DOHMH. Alternatively, sensitivity and specificity can be calculated using investigator review as gold standard denominators.

Many important research questions can be investigated by combining existing administrative, registry, and financial data sources. Although data linkages can be complex and resource-intensive activities, they may be less expensive and time-consuming, and more illuminating, than primary data collection such as population-representative surveys. Although each process will be unique based on the population, available identifiers, and purposes of the resulting data set, there are challenges common to them all. Successful methodologies should be shared among public health researchers to enhance understanding of data sets and the findings derived from them, as well as to provide greater potential for comparability between similar projects from different jurisdictions. As new technologies develop, publicizing strategies employed enhances public health research, resulting policy decisions, and, ultimately, individual health.

Table 1. Glossary of Terms Used to Describe the Longitudinal Study of Early Development Linkage Methods

Term	Definition
Linkage	The process of bringing records together to evaluate whether they represent to the same entity (e.g. same child or same mother) or the process of bringing data sets together for such an evaluation. Example: The comparison of records in the Birth data set with records in the Death data set is a linkage.
Match	Joined records that result from the linkage process. Example: If record A has been deemed to represent the same entity as record B, the pairing is a match. While the process is being developed and evaluated, the pairing is a potential match or potential match set.
Unresolved duplication	The occurrence of more than one identification number for the same unique entity within a data set. Example: If the records for identification number 1234 and identification number 5678 actually both belong to the same child, then the data set contains unresolved duplication.
Phase	A section of a linkage process in which the execution methods differ. Example: Our linkage of children was executed in two phases; the first involved only a subset of records and all potential matches were reviewed; the second involved all records and only a sample of potential matches were reviewed.
Blocking	The construction of smaller groups of records that all have identical values on the particular variables used to block. Only those within the same block will be further assessed as to whether they belong to the same entity. Example: The combination of code for child's first name, code for child's last name, child's full date of birth, and ZIP code was used as blocking criteria. Records that had the exact same values on all four variables were placed in the same group and compared.
Primary record	Within a match set or potential match set, one record is deemed the primary record. In QualityStage, the most complete record becomes the primary record. Example: Within potential match set 8765, record A is the primary record. It has values for child's first name, middle name, last name, date of birth, address, and mother's date of birth.
Secondary record	Within a match set or potential match set, the record(s) that represent the same child as the primary record are referred to as secondary records. These records are less complete than the primary record. Example: Within potential match set 8765, record B represents the same child as the primary record (record A), but is less complete. It has values for child's first name, last name, date of birth, and address but not child's middle name or mother's date of birth.
Weight	An indication of how similar one record is to another. A higher weight indicates that the pairing is more likely to truly represent the same entity than a pairing with a lower weight. Example: A record with a weight of 152 is more likely to truly represent the same entity as its primary record than a record with a weight of 74.

Term	Definition
Reviewer concordance	Agreement between the independent reviewers regarding whether or not the secondary record represented the same entity as the primary record. Example: If both reviewers determined that the primary and secondary records were the same child in 75 instances, both classified the secondary record as a different child in 20 instances, and the reviewers reached opposite conclusions in 5 instances, reviewer concordance is 95% [{(75+20)/(75+20+5)}*100 = 95%]
Threshold weight	The weight used to separate potential matches, without hand review, into those that are accepted and those that are rejected. Each pass will have its own threshold weight, although that value may be zero. Example: In pass seven, the threshold weight was 43; any potential match with a weight of 43 or higher was accepted whereas any potential match with a weight below 43 was rejected.
Non-match	Records that have been determined to represent more than one entity, either because they fell below the threshold weight or because they were not brought together during the linkage. Example: All potential matches from pass seven with a weight less than 43 were classified as non-matches.
False match	Records that have been classified as the same entity (i.e., a match) after the linkage is finalized but in actuality are not the same entity. Example: After the linkage was completed, a final sample of records was hand reviewed to determine how many children were actually a composite of more than one child, or how many were false matches.
Cumulative false match rate	The proportion of false matches, determined by hand review, among all of the potential matches with the same or higher weight. Example: Pass seven had a cumulative false match rate of 0.992 at weight 43 because there were 706 potential matches reviewed that had a weight of 43 or higher, 7 of which were false matches [(7/706)*100 = 0.992].
False non-matches	Records that were not joined as a result of the linkage process but in fact do belong to the same entity. Example: Some potential match sets with weights below the threshold weight actually represent one child. Since they are classified as non-matches because they fall below the threshold, they are false non-matches.
Invalid birth interval	Within a set of siblings, children with dates of birth more than two days but less than seven months apart. Example: If two children identified as siblings were born January 1, 1998 and May 1, 1998, this would represent an invalid birth interval within that sibling set.

Table 2. Characteristics of 5 New York City Data Sets Linked to Create the Longitudinal Study of Early Development Data Warehouse, January 1994–June 2007

	Data source				
Variable	Birth Certificate Registry	Early Intervention Administrative Data	Lead Poisoning Prevention Program Registry	Department of Education Administrative Data	Death Certificate Registry
Birth years included	1994 - 2004	1994 - 2004	1994 - 2004	1994 - 2001	1994 - 2004
Data through	1994 - 2004	1994 - 2004	1994 - 2007	1994 - June 2007	1994 - 2006
Number of observations	1,380,608	251,757	7,772,305	617,934	8,331
Number of children	1,380,608	156,834	1,469,265	617,934	8,331
Unique children	Yes	No	No	No	Yes
Identifiers					
Child's first name, last name, date of birth, sex, address	Yes	Yes	Yes	Yes	Yes
Child's middle name	Yes	Initial	Yes	Initial	Yes
Child's social security number	Yes	Yes	No	No	Yes
Mother's first and last name	Yes + Maiden	Yes	Yes	Guardian	No
Mother's date of birth	Yes	Yes	Yes	No	No
Father's last name	Yes	Yes	No	Guardian	Yes
Father's first name	Yes	Yes	No	Guardian	No

Table 3. Characteristics of Children With Records in the Longitudinal Study of Early Development Data Warehouse and Their Record Sources, New York City, January 1994 – June 2007

Characteristic	*n*	%
Child's sex		
Male	985,619	50.73
Female	934,050	48.07
Unknown	23,273	1.20
Child's race/ethnicity		
Hispanic	314,405	16.18
Black (non-Hispanic)	265,194	13.65
White (non-Hispanic)	144,041	7.41
Asian/Pacific Islander	88,751	4.57
Other race/ethnicity	7,062	0.36
Multiple race/ethnicity	7,474	0.38
Unknown race/ethnicity	1,116,015	57.44
Mother's race/ethnicity		
Hispanic	446,069	22.96
Black (non-Hispanic)	364,707	18.77
White (non-Hispanic)	407,281	20.96
Asian/Pacific Islander	154,104	7.93
Other race/ethnicity	4,311	0.22
Unknown race/ethnicity	566,470	29.16
Child's year of birth		
1994	216,924	11.16
1995	206,486	10.63
1996	195,293	10.05
1997	178,378	9.18
1998	176,682	9.09
1999	172,635	8.89
2000	171,958	8.85
2001	167,226	8.61
2002	156,717	8.07
2003	153,246	7.89
2004	147,397	7.59
Data sources		
Birth	1,380,535	71.05
Early Intervention	155,880	8.02
Lead	1,417,655	72.96
Education	609,321	31.36
Death	8,331	0.43
Number of data sources		
1	826,294	42.53
2	664,144	34.18
3	392,955	20.22
4	59,470	3.06
5	79	0.00

Table 4. Overlap of Data Sources for Children With Records in the Longitudinal Study of Early Development Data Warehouse, by Number of Sources, New York City, January 1994−June 2007

Sources	*n*	% Among all Children	% Among Children With More Than 1 Source
One source			
Birth	360,979	18.58	
EI	8,944	0.46	
Lead	387,898	19.96	
Education	67,416	3.47	
Death	1,057	0.05	
Two sources			
Birth and EI	12,272	0.63	1.10
Birth and Lead	496,265	25.54	44.44
Birth and Education	58,390	3.01	5.23
Birth and Death	5,703	0.29	0.51
EI and Lead	7,549	0.39	0.68
EI and Education	1,329	0.07	0.12
EI and Death	11	0.00	0.00
Lead and Education	82,587	4.25	7.40
Lead and Death	36	0.00	0.00
Education and Death	2	0.00	0.00
Three sources			
Birth, EI and Lead	51,580	2.65	4.62
Birth, EI and Education	8,476	0.44	0.76
Birth, EI and Death	643	0.03	0.06
Birth, Lead and Education	326,231	16.79	29.22
Birth, Lead and Death	444	0.02	0.04
Birth, Education and Death	11	0.00	0.00
EI, Lead and Education	5,557	0.29	0.50
EI, Lead and Death	8	0.00	0.00
EI, Education and Death	2	0.00	0.00
Lead, Education and Death	3	0.00	0.00
Four sources			
Birth, EI, Lead and Education	59,138	3.04	5.30
Birth, EI, Lead and Death	232	0.01	0.02
Birth, EI, Education and Death	52	0.00	0.00
Birth, Lead, Education and Death	40	0.00	0.00
EI, Lead, Education and Death	8	0.00	0.00
Five sources			
Birth, EI, Lead, Education and Death	79	0.00	0.01

Abbreviation: EI, Early Intervention.

Table 5. The Proportion of Records in Each Source That Contained Matches in the Corresponding Data Sources, Longitudinal Study of Early Development Data Warehouse, New York City, January 1994 - June 2007

	Data Source of Records That Matched (Numerator)				
	Birth	EI	Lead	Education	Death
Data Source of Records (Denominator)					
Birth	0[a]	9.60	67.65	44.79[b]	0.52
EI	84.98[a]	0.61[c]	79.65[a]	61.99[ab]	0.66
Lead	65.88[a]	8.76	3.44[c]	44.85[b]	0.06
Education	74.25[a]	12.25	77.73[d]	1.29[ab]	0.03
Death	86.47	12.42	10.20	2.96[b]	0.00

Abbreviation: EI, Early Intervention.
[a] Within the expected range of the linkage yield
[b] Children born after 2001 were excluded from the denominator since the Education data set only contained children born 1994−2001.
[c] Below the expected range of the linkage yield
[d] Above the expected range of the linkage yield

A. Records from Multiple Sources Combined Into 1 Data Set

Record Number	Child's first name	Child's last name	Child's date of birth	Child's address	Mother's date of birth
23	Melissa	Pfeiffer	11/12/2001	123 Main Street	01/09/1975
42	Melissa	Pfeiffer	11/12/2001	862 Pine Street	09/01/1975
57	Malisa	Pfeiffer	11/02/2001	862 Pine Street	09/01/1975
68	Melissa	Pfeiffer-Jones	12/11/2001	862 Pine Street	09/01/1975
81	Michelle	Pfeiffer	01/28/2001	123 Main Street	01/09/1975

B. Pass 1 Linkage Criteria, Blocks, and Matches. Block on child's last name, child's month of birth, child's year of birth. Additional linkage variables: child's first initial, child's day of birth, child's address, mother's date of birth.

Record Number	Child's first name	Child's last name	Child's date of birth	Child's address	Mother's date of birth	Block
23	Melissa	Pfeiffer	11/12/2001	123 Main Street	01/09/1975	First
42	Melissa	Pfeiffer	11/12/2001	862 Pine Street	09/01/1975	First
57	Malisa	Pfeiffer	11/02/2001	862 Pine Street	09/01/1975	First
68	Melissa	Pfeiffer-Jones	12/11/2001	862 Pine Street	09/01/1975	Second
81	Michelle	Pfeiffer	01/28/2001	123 Main Street	01/09/1975	Third

C. Pass 2 Linkage Criteria, Blocks, and Matches. Block on child's first name, mother's date of birth. Additional linkage variables: child's day and month of birth (transposed), child's address.

Record Number	Child's first name	Child's last name	Child's date of birth	Child's address	Mother's date of birth	Block
23	Melissa	Pfeiffer	11/12/2001	123 Main Street	01/09/1975	First
42	Melissa	Pfeiffer	11/12/2001	862 Pine Street	09/01/1975	Second
57	Malisa	Pfeiffer	11/02/2001	862 Pine Street	09/01/1975	Third
68	Melissa	Pfeiffer-Jones	12/11/2001	862 Pine Street	09/01/1975	Second
81	Michelle	Pfeiffer	01/28/2001	123 Main Street	01/09/1975	Fourth

D. Final Outcome of Child Linkage, With Connections Between Passes Made.

New ID number	Original record number	Child's first name	Child's last name	Child's date of birth	Child's address	Mother's date of birth
1	23	Melissa	Pfeiffer	11/12/2001	123 Main Street	01/09/1975
1	42	Melissa	Pfeiffer	11/12/2001	862 Pine Street	09/01/1975
1	57	Malisa	Pfeiffer	11/02/2001	862 Pine Street	09/01/1975
1	68	Melissa	Pfeiffer-Jones	12/11/2001	862 Pine Street	09/01/1975
2	81	Michelle	Pfeiffer	01/28/2001	123 Main Street	01/09/1975

Figure 1. Schematic of Longitudinal Study of Early Development Child Linkage Process With Hypothetical Data, New York City, January 1994—June 2007.

References

1. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu. Rev. Public Health.* 2011;32:91-108.
2. Papadouka V, Schaeffer P, Metroka A, et al. Integrating the New York Citywide Immunization Registry and the Childhood Blood Lead Registry. *J Public Health Management Practice.* 2004;10(Suppl):S72-S80.
3. Jaro MA. Probabilistic linkage of large public health data files. *Stat Med.* 1995;14(5-7):491-498.
4. Howe GR. Use of computerized record linkage in cohort studies. *Epidemiol Rev.* 1998;20(1):112-121.
5. Tuoto T, Cibella N, Fortini M, et al. RELAIS: Don't get lost in a record linkage project. In Proceedings of the Federal Committee on Statistical Methodologies (FCSM) Research Conference, Arlington VA, USA (2007). (http://www.fcsm.gov/07papers/Tuoto.VI-C.pdf). (Accessed April 13, 2011).