Appendix: Data Linkage Method with Deterministic Algorithm.

We used a deterministic data linkage method described by Martin7 and adapted it to link records in the STD syphilis surveillance database (STD database) and the California Birth Statistical Master File (BSMF). For the purposes of this appendix, names were fabricated to protect the identity of women.

*Data cleaning and standardization.* Before the linkage, variables were cleaned as described by Martin.7 Punctuation was removed, text was made upper case and common misspellings were corrected. We modified this approach to account for middle names (Step 1) and then created a wide dataset with separate name variables for all names containing more than one part (Step 2). For first names with more than one part in Step 1, both first and last name variables were generated for additional parts.

**Original STD Data Structure**

|  |  |  |  |
| --- | --- | --- | --- |
| **STD\_ID** | **MOM\_FIRST** | **MOM\_MIDDLE** | **MOM\_LAST** |
| 1 | TRACY | LEE | ATKIN |

**Step 1: Data Cleaning Results**

|  |  |  |
| --- | --- | --- |
| **STD\_ID** | **MOM\_FIRST** | **MOM\_LAST** |
| 1 | TRACY LEE | ATKIN |

**Step 2: Wide Data Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **STD\_ID** | **MOM\_FIRST1** | **MOM\_FIRST2** | **MOM\_FIRST3** | **MOM\_LAST1** | **MOM\_LAST2** | **MOM\_LAST3** |
| 1 | TRACY | LEE | TRACYLEE | LEE | ATKIN | LEEATKIN |

We then interleaved the first and last names to create one long dataset with all possible combinations of first and last names, which omitted first and last name combinations in which the same name would have been used in both (Step 3).

**Step 3: Long Data Results**

|  |  |  |
| --- | --- | --- |
| **STD\_ID** | **STD\_FIRST** | **STD\_LAST** |
| 1 | TRACY | LEE |
| 1 | TRACY | ATKIN |
| 1 | LEE | ATKIN |
| 1 | TRACY | LEEATKIN |
| 1 | TRACYLEE | ATKIN |

In the final step, we interleaved the last names of reported sexual partners (Step 4). In addition, names were truncated to 8 characters in the 2013 BSMF, we inserted all possible combinations of first and last names truncated to 8 characters after deleting duplicate combinations (not shown).

**Step 4: Final STD Data Structure**

|  |  |  |
| --- | --- | --- |
| **STD\_ID** | **STD\_FIRST** | **STD\_LAST** |
| 1 | TRACY | LEE |
| 1 | TRACY | ATKIN |
| 1 | LEE | ATKIN |
| 1 | TRACY | LEEATKIN |
| 1 | TRACYLEE | ATKIN |
| 1 | TRACY | PARKER |
| 1 | LEE | PARKER |
| 1 | TRACY | LEEPARKER |
| 1 | TRACY | ATKINPARKER |
| 1 | LEE | ATKINPARKER |
| 1 | TRACY | LEEATKINPARKER |
| 1 | TRACYLEE | ATKINPARKER |

The same processing was performed on BSMF data except infants’ and fathers’ last names were interleaved in Step 4.

*Blocking phase.* To reduce processing time, we identified a subset of potential matches during an initial blocking phase. For non-CS mothers who had multiple births during the time period of the study, we restricted potential matches to pregnancies and deliveries during which mothers were diagnosed with syphilis. For CS mothers who had multiple births during the time period of the study, we restricted potential matches to pregnancies and deliveries during which their infant was reported with CS in the STD database.

Records were included in the subset of potential matches if they agreed on any of the following:

* First name and last name
* First name and date of birth
* Last name and date of birth
* Date of birth and county of residence

*Application of deterministic algorithm.* Records in the STD database were linked to the BSMF using a deterministic algorithm. The algorithm ensured that all records of mothers who had an infant with CS in the STD database linked to records in the BSMF. We adapted Martin’s code7 to account for name-related data entry errors, including misspellings and transpositions. We also adapted Mullins’ code20 to generate standardized street addresses.

Records were declared a match if they agreed on date of birth, first name, and one of the following:

* Street address
* Last name and county of residence
* Last name and delivery date ± 1 day (for CS mothers) or delivery date within 17 weeks before 2 weeks after estimated due date (for non-CS mothers)

If records did not agree on date of birth, they were declared a match if they agreed on one of the following:

* First name, last name, day and month of birth, three of four digits of year of birth, county
* First name, last name, year of birth, transposed day and month of birth, county
* First name, last name, year of birth, day or month of birth, street address

If records were deidentified in the BSMF (no name or address information available), records were declared a match if they agreed on all of the following:

* Date of birth
* County of residence
* Delivery date ± 1 day (for CS mothers) or delivery date within 17 weeks before 2 weeks after estimated due date (for non-CS mothers)

For each pair of records, match or nonmatch status was determined using the rules above, and match markers were generated indicating agreement or disagreement on each individual identifier (date of birth, first name, last name, delivery date, county of residence, and street address). Duplicate matches were resolved by comparing individual identifiers and keeping the match that linked on the greatest number of identifiers.

*Validation of linkage results.* We reviewed the individual identifiers of all matches to validate the results.