Probabilistic data linkage: generating a reproductive histories dataset from states' vital records data

by

Basma Nihad Dib

Bachelor of Medicine and Surgery, Jordan University of Science and Technology, 2015

Submitted to the Graduate Faculty of the

Department of Epidemiology

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Public Health

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This essay is submitted

by

Basma Nihad Dib

on

December 17, 2021

and approved by

Essay Advisor: Lisa M. Bodnar
PhD, MPH, RD, Professor, Epidemiology, Graduate School of Public Health, University of Pittsburgh

Ada O Youk
PhD, Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Sara M. Parisi
MS, MPH, Research Data Analyst, Epidemiology, Graduate School of Public Health, University of Pittsburgh

Probabilistic data linkage: generating a reproductive histories dataset from states' vital records data

Basma Nihad Dib, MPH

University of Pittsburgh, 2021

Abstract

Datasets that follow pregnancy histories over time are lacking. In this pilot study, we used Pennsylvania's fetal death and birth records to generate a longitudinal maternal dataset by linking records for the same mother to each other. We explored how to best achieve this linkage when lacking a unique record identifier. We demonstrated how Stata's existing probabilistic matching tools can use nonunique identifiers to facilitate record linkage. To validate the effectiveness of this probabilistic linkage, we compared its results to the linkage results generated from deterministically linking the records using social security numbers. Compared to the deterministic linkage, the probabilistic linkage had a sensitivity of 94.3% and a positive predictive value of 96.7%. Our pilot study can serve as a guide for researchers in other states to generate longitudinal maternal datasets from their states' vital records. Such longitudinal datasets can be a valuable resource for conducting epidemiologic analyses in the field of maternal and child health and answering research questions that relate to the period between pregnancies. Results from these studies can be used to improve health outcomes of mothers and children.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Overview

Longitudinal datasets that follow pregnancy histories over time can be a valuable resource for conducting epidemiologic analyses. They allow for studying many causal questions, including those that relate to the period between pregnancies. State health departments in the United States collect data on fetal deaths and births as part of each state's vital records system. These records have a large number of important variables related to characteristics of the mother and her fetus or newborn, as well as pregnancy and birth outcomes. In these records, each fetal death or birth is listed as a separate, unlinked record. Vital records rarely include a unique identifier that would allow linkage of these records over time. The primary goal of our work is to conduct a pilot study to explore how to best link a mother's pregnancy records to each other when lacking a unique identifier. This would generate a longitudinal dataset that follows mothers' pregnancy histories over time.

1.2 Record Linkage

Record linkage is a powerful tool in the field of public health. Different record linkage techniques have been widely used to link data from two or more datasets. Linkage of data combines different variables relating to the same record, and its techniques rely on the availability of personal identifying variables. Ideally, two datasets can be easily linked to each other if a unique personal identifier exists in both datasets. Using this exact matching technique is known as deterministic matching. As noted above, however, a universal unique personal identifier often does not exist. In this case, a number of nonunique personal identifiers can be used to link individuals using a technique called probabilistic matching, also referred to as fuzzy matching. Probabilistic matching has been used to link vital records to hospital discharge records,[1] link trauma registries to traumatic Brain Injury Model Systems,[2,3] and create a set of maternally linked sibships,[4] to name a few.

1.3 Probabilistic Matching

When linking two datasets using probabilistic matching, each record in the first dataset is compared with records in the second dataset based on a set of nonunique personal identifying variables available in both datasets, such as date of birth, first name, and last name. Each identifying variable used in the matching process is assigned a match weight and a nonmatch weight. These weights are used to measure the likelihood that two records are for the same person.

The values of the match weights, commonly referred to as agreement weights, reflect the relative likelihood that a variable match indicates a true records match. Match weights correspond to how likely the values of the variable are to be repeated in the dataset (i.e., variables with more unique values are assigned higher match weights). For instance, social security number will have a very high match weight, as duplicates are not expected across records (unless being for the same individual or due to errors). In contrast, a categorical variable, like maternal race, will have a relatively low match weight, as many duplicates are expected across records.

On the other hand, nonmatch weights, commonly referred to as disagreement weights, reflect the relative likelihood that a mismatch on a variable indicates that the records truly do not match. A small value for a nonmatch weight indicates that mismatches are expected even if the records match. For instance, maternal first name will have a higher nonmatch weight than maternal last name. This is because two records that do not match on the maternal first name are unlikely to be a match as very few changes in the first name are expected over time. In contrast, two records that do not match on the maternal last name have a higher chance of being a match, as changes to the maternal last name occur more commonly than changes to the maternal first name over time.

Each linked record pair is assigned an overall match score. There are several steps to this process. First, all identifying variables are compared between the two records. If the values of a variable are the same in both records, the record pair is given the match weight for that variable. Otherwise, if the values are not the same, the record pair is given the nonmatch weight for that variable. Finally, the weights of all identifying variables are summed to generate an overall match score for the record pair. Researchers need to determine a cut-point for match scores; record pairs with an overall match score at or above the cut-point are considered to be true matches and those below are not considered to be true matches. The probabilistic matching process should be followed by validation of the generated matches. Manually reviewing a random sample of the generated matches is one way to validate the success of the matching process, however, it is tedious and not efficient.

1.4 Probabilistic Matching Techniques in Prior Research

Looking at previous research, subtle methodological differences exist in applying the concepts of probabilistic matching. Here, we highlight the main concepts of probabilistic matching in light of previous research:

1.4.1 Preprocessing

Record linkage is preceded by preprocessing of the data. This ensures that the identifying variables used for the linkage exist appropriately in the two datasets and have the same formats. Preprocessing includes parsing a field into relevant subcomponents, standardizing character strings, and removing extra spaces.[5] For instance, we could parse the mother's name variable into first, middle, and last names variables, and then convert all their values into uppercase. Preprocessing helps to achieve higher quality matches from the linkage.

1.4.2 Estimating match and nonmatch weights

Estimating each variable's match and nonmatch weights is the cornerstone of probabilistic matching. Estimating these weights depends on measuring two quantities:

1. m-probability: the probability that a pair of records agree for a certain linkage variable, given that both records belong to the same individual.
2. u-probability: the probability that a pair of records agree for a certain linkage variable, given that both records belong to different individuals.

m- and u- probabilities, by definition, require knowing the true match status for each linked record pair. However, the true status is usually not known. Therefore, m- and u-probabilities are frequently estimated relying on the best linking variable available in the dataset. Match weights and nonmatch weights are, then, estimated as follows:

- Match weight = m/u
- Nonmatch weight = (1-m)/(1-u)

To simplify computations, the $\log_2$ transformation of these ratios is used in practice.[2,6,7]

1.4.3 Blocking

When applying probabilistic matching to link two datasets, the first record in the first dataset is compared to each record in the second dataset, then the second record in the first dataset is compared to each record in the second dataset, and so on until the last record in the first dataset. When attempting to link two large datasets, such as vital records, one can imagine that a very large number of comparisons will take place. Furthermore, most of these comparisons are highly unlikely to be true matches. So, this can be extremely computationally intensive and requires long runtimes.

To limit the number of comparisons needed to link two large datasets, one can "block" on certain variables.[3] Blocking on variables means that comparisons between the two datasets will happen only among records that have the same exact value for one or all of the blocking variables. For instance, when blocking on maternal date of birth, the comparisons will be made only among records that have exactly the same date of birth. It is important to note that multiple blocking variables can be used in an "OR" or "AND" fashion. For instance, when blocking on first name AND last name, comparisons will happen only among records that have exactly the same first and last names. Alternatively, when blocking on first name OR last name, comparisons will be among records that have exactly the same first name or exactly the same last name. When expecting input errors or misspellings in the values of the blocking variables, it is better to opt for the "OR" blocking option.

1.4.4 Reporting results

Results of probabilistic record linkage are frequently reported in terms of sensitivity and positive predictive value.[8] Sensitivity, also known as true positive rate, estimates the percentage of true matches that the probabilistic matching process is able to identify. Positive predictive value estimates the percentage of matches generated from the linkage process that are, in reality, true matches. In order to estimate these values, researchers need to identify a “gold standard” which is basically the true match status. This can be conceptualized in many ways depending on the available data. Results generated using a set of the strongest personal identifiers, matched deterministically or probabilistically, can be regarded as the “gold standard”.[7] The probabilistic linkage using a less robust (but more readily available) set of identifying variables can be compared to this gold standard.

1.5 Using Stata Software for Record Linkage

In our work, we chose to use Stata software to facilitate our probabilistic record linkage. The package `dtalink` was developed in Stata to link large data files, and within this package is an option `calcweights` that allows for estimating match and nonmatch weights.[9] As noted above, weight estimation relies on knowing the true match status for each linked record pair. Record pairs that match on the most unique identifying variable can be assumed to be true matches. Using `dtalink`, users can give the most unique identifying variable a large match weight and set the weights for the remaining identifying variables to zero. Users need to choose an overall match score cut-point, so that record pairs with match scores at or above the cut-point are considered true matches. For the purpose of weights estimation, users should set the overall match score cut-point at the value of the match weight given for the most unique variable. The `dtalink` command with the `calcweights` option can then be executed on the dataset. This option tracks the number of times a variable matched among matched pairs and among nonmatched pairs. It, then, uses these percentages to compute estimated weights for the remaining identifying variables. Weights are calculated as follows:

- Match weight = $\log_2(p_1/p_2)$
- Nonmatch weight = $\log_2((1-p_1)/(1-p_2))$

p_1 is the percentage of times the variable matched among matches, and p_2 is the percentage of times the variable matched among nonmatches.[9] These calculations correspond to the match and nonmatch weights estimations mentioned earlier. We believe the `dtalink` command with the `calcweights` option is quite useful to estimate weights for the identifying variables used in probabilistic matching.

Another Stata command, `reclink2,` also facilitates probabilistic matching. Within this command are a number of user-friendly options that allow for reliable and efficient data linkage. First, match and nonmatch weights estimated using the `dtalink` command can be imputed in the `reclink2` command using the `wmatch` and `wnomatch` options. The `orblock` options allows for blocking on a set of variables in "OR" fashion, and the `required` option allows for blocking in "AND" fashion. Also, the `minbigram` option specifies the minimum bigram value to declare two string values as matched. Bigram is an approximate string comparator that can efficiently account for transpositions of characters within a string. It gives a value from zero to one for any two linked strings rather than declaring them as exact or not exact matches. Furthermore, the `manytoone` option allows records from the second dataset to be matched to multiple records from the first dataset. Moreover, the `npairs(#)` option specifies the number of matched pairs that the program will retain above the minimum score threshold. Finally, when executing the `reclink2` command, each linked record pair is given an overall match score from zero to one. An overall match score of one is a perfect match.[5]

While the `dtalink` command has been developed to facilitate probabilistic matching, its basic technique declares two string values as a match only if they are exactly the same value. Assigning a partial weight for strings that are slightly different requires using advanced coding methods.[9] In light of these limitations, we believe that the `reclink2` command is a more convenient tool to facilitate probabilistic linkage of our data.

1.6 State of Pennsylvania Fetal Death and Birth Records

The purpose of our work is to link mothers' pregnancy histories over time and generate a longitudinal maternal dataset using the State of Pennsylvania fetal death and birth records. A number of technical issues that impact the linkage process exist in this dataset. First, the social security number is the only unique maternal identifier in this dataset, but it is not routinely provided with the state's records and only available for a small portion of the years of data. However, a number of nonunique maternal identifiers are present, such as date of birth, first name, middle name, last name, height, race, and zip-code of residence, as well as father's first and last names. Other technical issues are errors in entering the social security numbers and dates of births, misspellings of names, last names and zip-codes that have changed over time, and inaccurate measurements of the mothers' heights.

1.7 Gaps in Knowledge

Longitudinal datasets that follow mothers' pregnancy histories over time are lacking. Though probabilistic matching is not a novel technique per se, its application in generating a longitudinal maternal dataset from the states' fetal death and birth records is novel. Moreover, demonstrating how Stata's existing packages facilitate probabilistic matching and validating how well they work adds to the literature on use of probabilistic matching in practice.

1.8 Public Health Significance

Generating a longitudinal maternal dataset would provide a rich data source to conduct epidemiologic analyses in the field of maternal and child health. It would provide the opportunity for studying a wide range of research questions requiring longitudinal data. Results from these studies can be used to improve health outcomes of mothers and children. Moreover, researchers in other states can use the results of our pilot study as a guide to generate longitudinal maternal datasets from their states' vital records.

2.0 Objectives

The specific objective of our pilot study is to investigate the effectiveness of Stata's probabilistic matching tools for linking maternal records from Pennsylvania's fetal death and birth records (2003-2007) when a unique identifier is not present. To achieve this, we will compare the probabilistic linkage that utilizes a set of nonunique identifying variables not including the social security numbers to the deterministic linkage that utilizes the social security numbers ("the gold standard"). Throughout our work, we will illustrate how existing packages in Stata can be used to facilitate the probabilistic linkage process.

3.0 Methods

3.1 Preprocessing

First, we prepared our data for linkage. We recoded missing data values of the social security number with appropriate missing values in Stata. We removed extra spaces before, after, and within variable values and converted all string values into uppercase. We checked the percentage of missing values for all of the identifying variables used in the matching process using Stata's `mdesc` command.

3.2 The Basic Methodology of our Linkage

The number of births and fetal deaths in Pennsylvania from 2003 to 2007 was 741,282. We dropped the records that were missing social security number (n = 29,331). Record linkage requires having two datasets to link to each other. To achieve this, we subset the first six-month interval of the dataset and probabilistically linked it to the rest of the dataset. We chose to subset a six-month interval as we assumed that it is very unlikely that mothers had two births and/or fetal deaths within a six-month period. We will refer to the first dataset as the "master" dataset and the second dataset as the "using" dataset. For example, the first master dataset was births and fetal deaths from January 1st, 2003 to June 30th, 2003. And its corresponding using dataset was births and fetal deaths from July 1st, 2003 to Dec 31st, 2007. By probabilistically linking these two datasets, records of mothers who experienced fetal deaths or births within the first six-month interval are matched to the same mothers' records of fetal deaths or births within subsequent years.

Once the first round of matching was completed, we removed the matched records from the dataset and set them aside. We, then, subset the next six-month interval of the dataset (July 1st, 2003 – Dec 31st, 2003) and linked it to the rest of the dataset (Jan 1st, 2004 – Dec 31st, 2007). Similarly, we removed the matched records from the dataset and appended them to the previously matched records. We repeated this probabilistic linkage process seven more times until the last six-month interval of the dataset.

3.3 Identifying the Best Gold Standard

To validate the effectiveness of our probabilistic record linkage, we needed to identify the best gold standard, which is a set of identifying variables that includes social security number and generates matches with the highest sensitivity and specificity. As we are aware that data entry errors affect the accuracy of social security numbers, we explored two different sets of variables as options for the gold standard.

First, we used the social security number alone to deterministically match the first master and using datasets. Second, we additionally dropped the records that were missing the mother's first name from the dataset (n = 817), and then we used the social security number and the first three letters of the mother's first name to deterministically match the first master and using datasets. We then compared the number of matches generated from the two linkages and manually reviewed the matches generated only by the first linkage but not by the second. We marked these manually reviewed matches as either true matches (were not generated by the second linkage due to first name misspellings or changes over time) or false matches (were generated by the first linkage due to errors in social security numbers). Based on the results of this manual review, we identified the best gold standard.

3.4 Probabilistic Matching

The nonunique maternal identifying variables we used for probabilistic matching were maternal date of birth, first name, middle name, last name, height, residence zip-code, as well as father's first and last names. Probabilistic matching of our data was carried out as follows:

3.4.1 Estimating match and nonmatch weights

We used Stata's `calcweights` option within `dtalink` command to estimate match and nonmatch weights for the identifying variables. To estimate the weights, we chose to rely on the social security number as it is the most unique identifier in our dataset. Then, we executed `dtalink` on our first master and using datasets while giving the social security number a match score of 20 and setting all match and nonmatch weights for other identifying variables at zero. We set the overall match score cut-point for declaring two records as matched at 20. Basically, this command tracks the number of times each identifying variable matched among pairs with exactly the same social security number and among pairs with different social security numbers. It, then, uses these numbers to compute estimated match and nonmatch weights for all identifying variables.

3.4.2 Data linkage

We used the `reclink2` command to link each 6-month interval master dataset to its corresponding using dataset using the following options:

- blocking on the mother's date of birth, first name, and last names in "OR" fashion
- minimum overall match score set at 0.80 for considering any linked record pair as a match
- minimum string bigram comparator set at 0.80 for considering any two string values as matched
- many-to-one matching to allow for matching all the records of the same mother that exist in the master dataset (i.e., twins) to their corresponding records in the using dataset
- maximum number of matched pairs that the program should retain set at six. We assumed that it is very unlikely for a mother to have more than six births and fetal deaths within a five-year period (2003-2007). This number was reduced according to the number of years between the master dataset and its corresponding using dataset.

3.4.3 Determining the appropriate overall match score cut-point

To start, we set our minimum overall match score at 0.80 for considering any linked record pair as a match. However, we wanted to identify a more precise overall match score cut-point for declaring matched pairs as true matches. To achieve this, we deterministically linked each master dataset to its corresponding using dataset using the gold standard. We then compared the matches generated probabilistically by executing `reclink2` with those generated by the deterministic matching. We identified the matches that were generated by both probabilistic and deterministic matching, and the matches that were generated only by probabilistic matching. We plotted the overall match scores of these matches and identified the appropriate overall match score cut-point (Figure 2).

3.4.4 Validating the linkage

After we had probabilistically linked each master dataset to its corresponding using dataset, and settled on an appropriate match cut-point, we validated the effectiveness of this linkage. We identified the number of matches generated by both probabilistic and deterministic linkage as well as the number of matches generated by one of the linkages but not the other. We reported these results in terms of sensitivity and positive predictive value.

4.0 Results

4.1 Missing Data

The number of births and fetal deaths in Pennsylvania from 2003 until 2007 was 741,282. Table 1 shows the number and percentage of missing values for the identifying variables used in the linkage.

Table 1 Number and percentage of missing values for the identifying variables used in the linkage

Variable	Number of missing values (N = 741,282)	Percentage of missing values (%)
Mother's social security number	29,331	3.96
Mother's date of birth	1,013	0.14
Mother's first name	1,019	0.14
Mother's middle name	104,025	14.03
Mother's last name	1,019	0.14
Father's first name	72,059	9.72
Father's last name	73,578	9.93
Mother's height	21,791	2.94
Mother's residence zip-code	1,593	0.21

4.2 Social Security Number as the Best Gold Standard

After dropping records that were missing social security number, we deterministically linked the first master and using datasets using the social security number alone. After additionally dropping records that were missing mother's first name, we deterministically linked the first master and using datasets using the social security number and the first three letter of the mother's first name. Table 2 shows the results of the two linkages.

Using the social security number alone for the first linkage generated 32,685 matches, while using both the social security number and the first three letters of the mother's first name for the second linkage generated 32,350 matches. We manually reviewed the 335 matches that were generated only from the first linkage but not from the second linkage. 215 (64%) of them were true matches that were not generated by the second linkage due to first name misspellings or changes over time, while 120 (36%) of them were false matches that were generated by the first linkage due to errors in social security numbers. As the false matches generated by the first linkage are only 120 of 32,685 total matches (0.4%), we chose to consider the social security number alone as the best gold standard.

Table 2 Deterministic linkages of the first master and using datasets to identify the best gold standard

	First linkage	**Second linkage**
Identifying variables used	Social security number	Social security number and first three letters of mother's first name
Number of records in master dataset*	69,593	69,593
Number of records in using dataset^	642,358	641,541
Number of generated matches	32,685	32,350

*Records from Jan 1st, 2003 until Jun 30th, 2003
^Records from Jul 1st, 2003 until Dec 31st, 2007

4.3 Probabilistic Matching

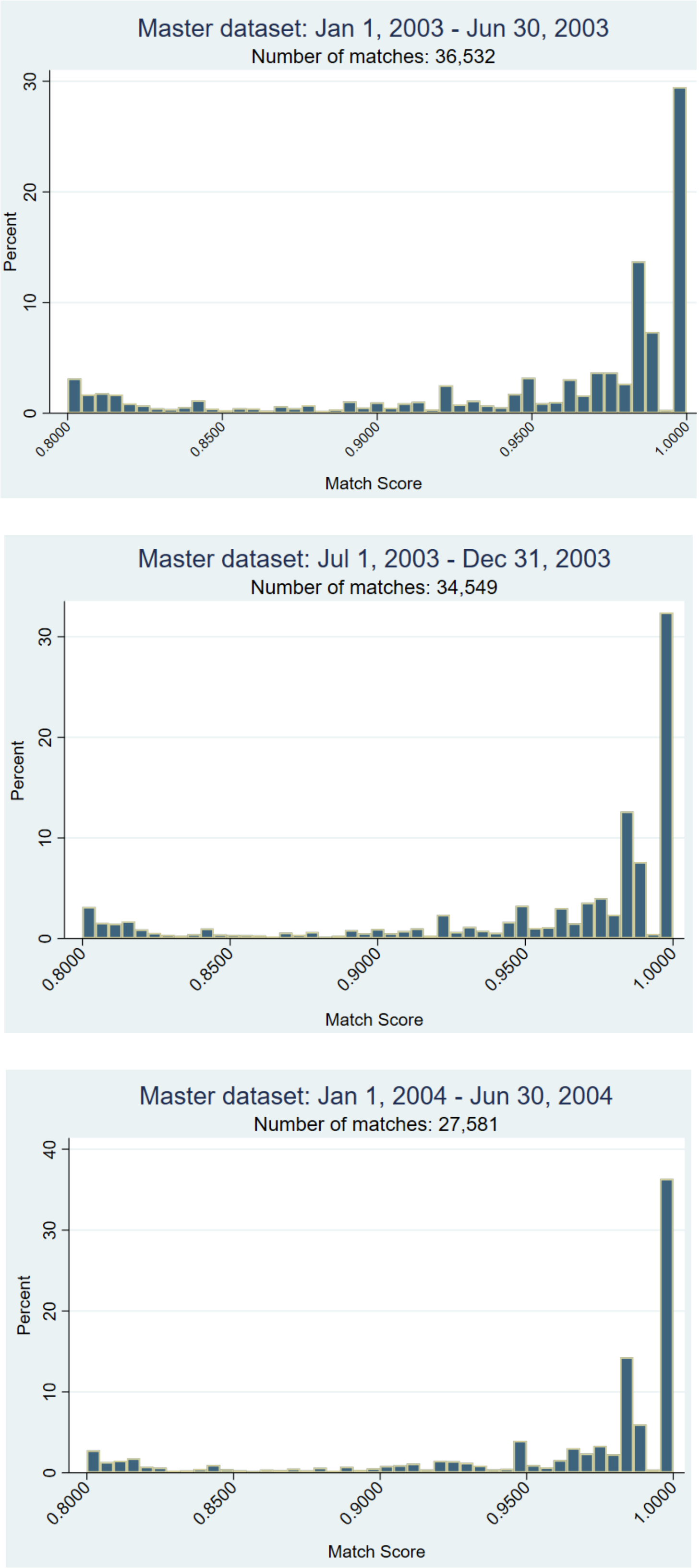
Results from executing `dtalink` command with `calcweights` option on the first master and using datasets are shown in Table 3. The highest match weight was for the date of birth, as it has the most unique values and the fewest number of duplicates across records. The lowest match weight was for the mother's residence zip-code, as it has the highest number of duplicates across records. The highest nonmatch weights were for the date of birth and mother's first name, as record pairs that do not match on these variables are unlikely to be for the same mother. We used these estimated weights throughout the probabilistic record linkage.

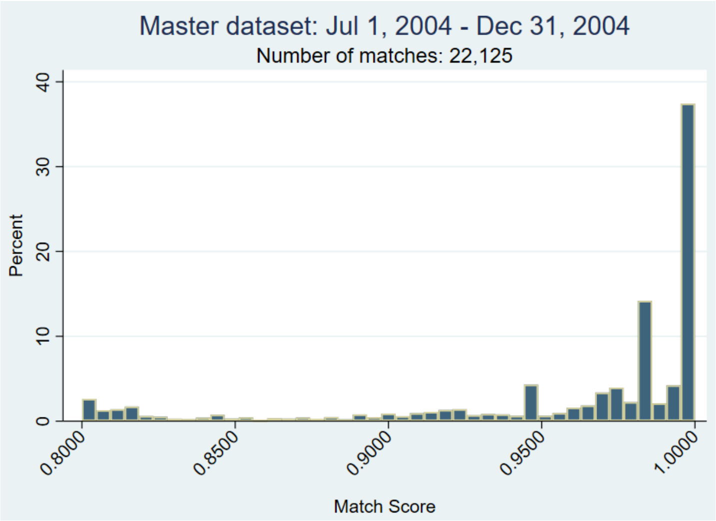
Table 3 Match and nonmatch weights estimated using the social security number as the gold standard identifier (weights are rounded to the nearest integer)

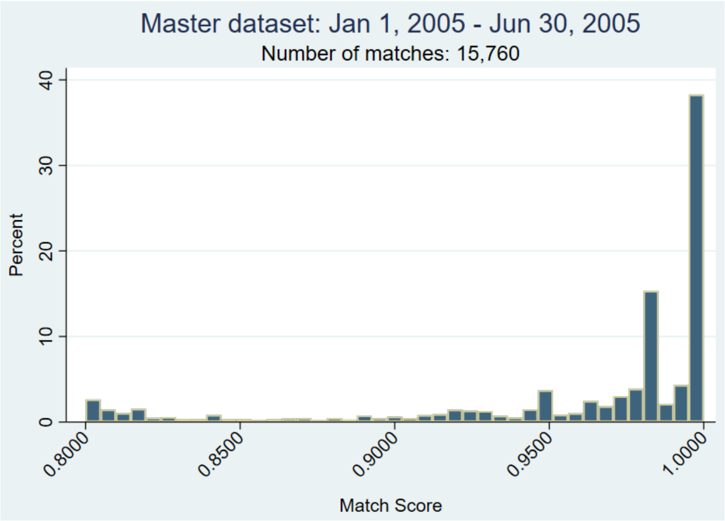
Identifying variable	Match weight	Nonmatch weight
Mother's date of birth	13	5
Mother's first name	8	5
Mother's middle name	5	2
Mother's last name	11	3
Father's first name	6	2
Father's last name	11	2
Mother's height	8	1
Mother's residence zip-code	3	1

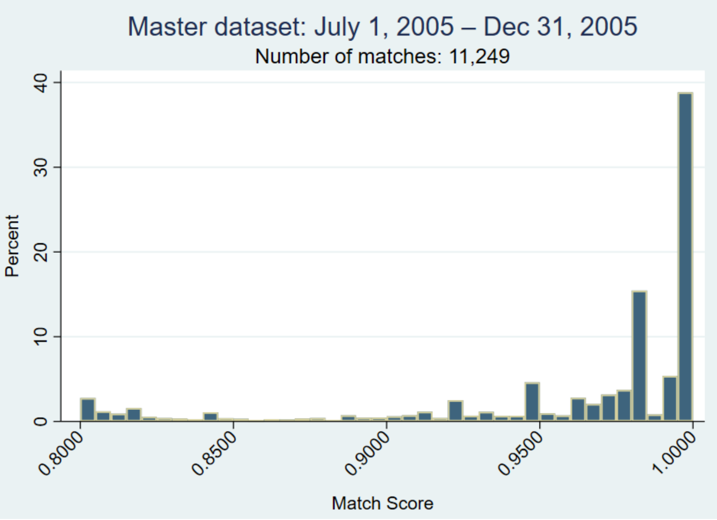
We executed `reclink2` on each master dataset to link it to its corresponding using dataset using a minimum overall match score of 0.8000 to start. Figure 1 shows the number of matches and distribution of the overall match scores for all the linkages. Except for the last linkage, the majority of the generated matches had match scores ≥ 0.95. The last linkage is linking records from Jan 1st, 2007 – June 30th, 2007 to records from Jul 1st, 2007 – Dec 31st, 2007. From this last linkage, the majority of the generated matches had match scores below 0.95, as relatively few births and fetal deaths for the same mother are expected to occur within one year.

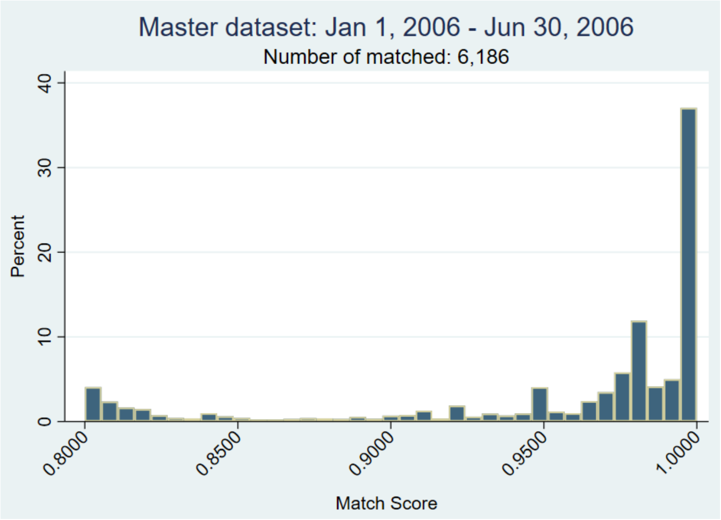
Figure 1 Histograms showing the distribution of the overall match scores for the matches generated from each probabilistic linkage

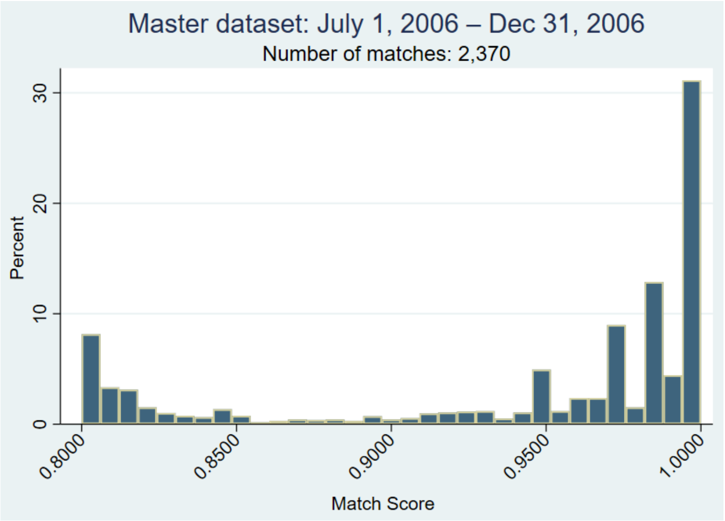


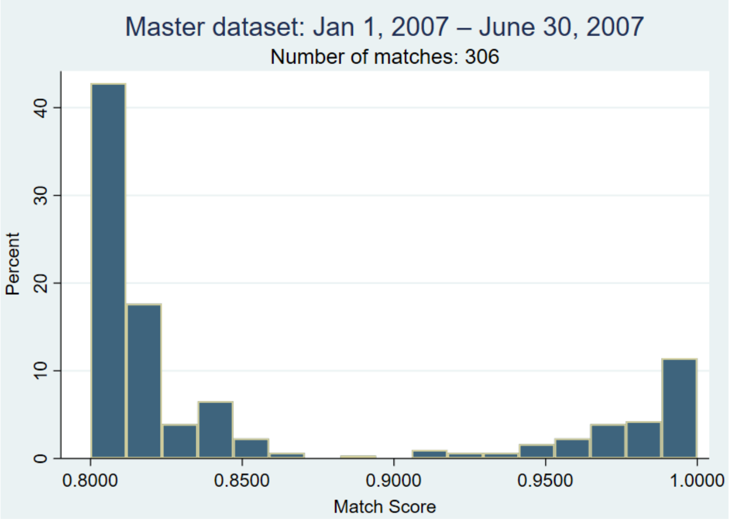

Master dataset: Jul 1, 2004 - Dec 31, 2004
Number of matches: 22,125
Percent
0
10
20
30
40
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score


Master dataset: Jan 1, 2005 - Jun 30, 2005
Number of matches: 15,760
Percent
0
10
20
30
40
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score


Master dataset: July 1, 2005 – Dec 31, 2005
Number of matches: 11,249
Percent
0
10
20
30
40
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score

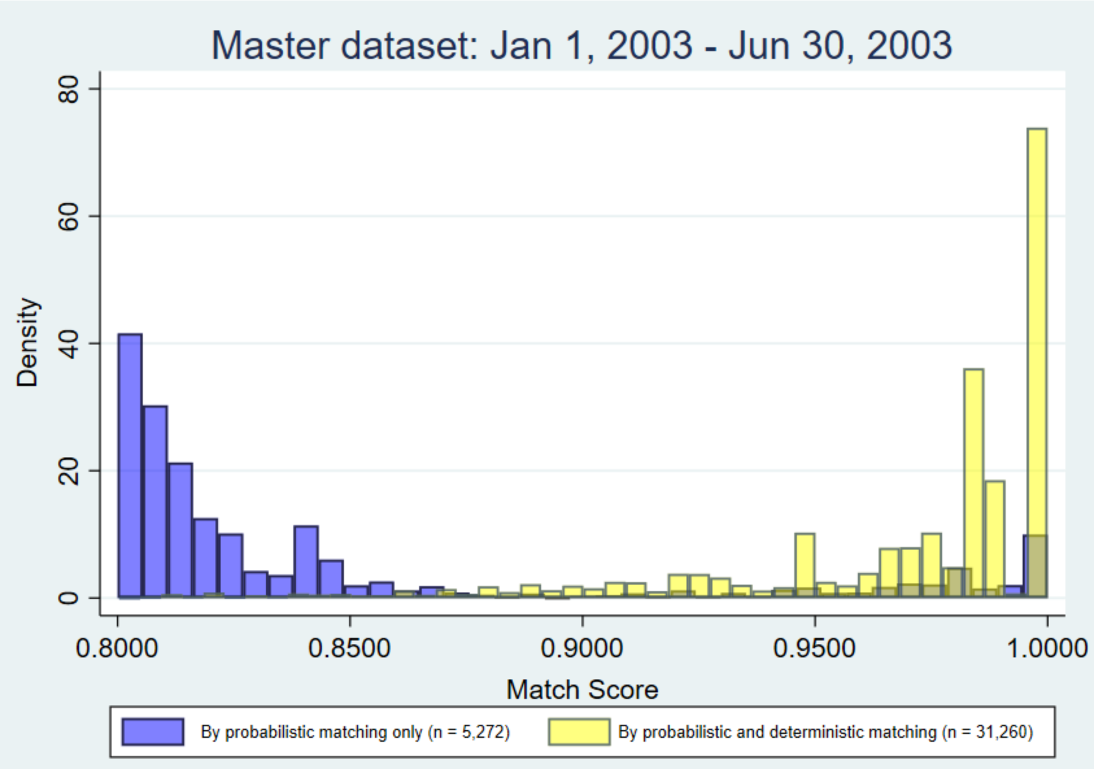

Master dataset: Jan 1, 2006 - Jun 30, 2006
Number of matched: 6,186
Percent
0
10
20
30
40
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score

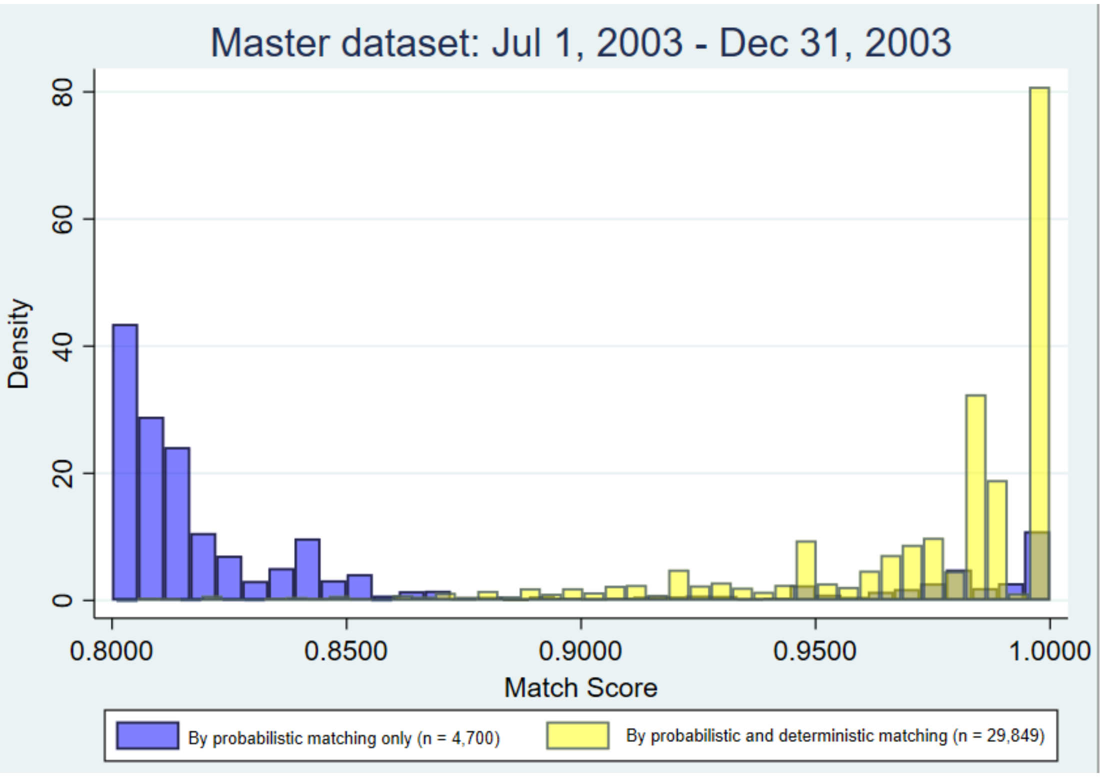

Master dataset: July 1, 2006 – Dec 31, 2006
Number of matches: 2,370
Percent
0
10
20
30
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score

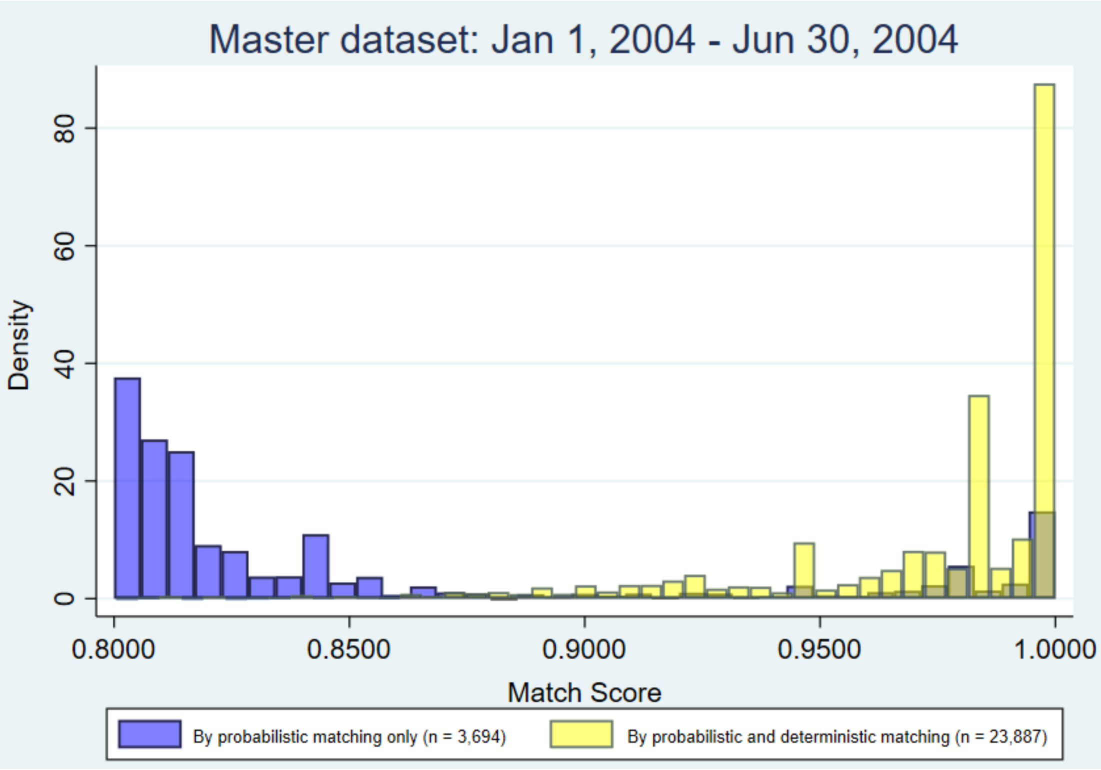

Master dataset: Jan 1, 2007 – June 30, 2007
Number of matches: 306
Percent
0
10
20
30
40
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score

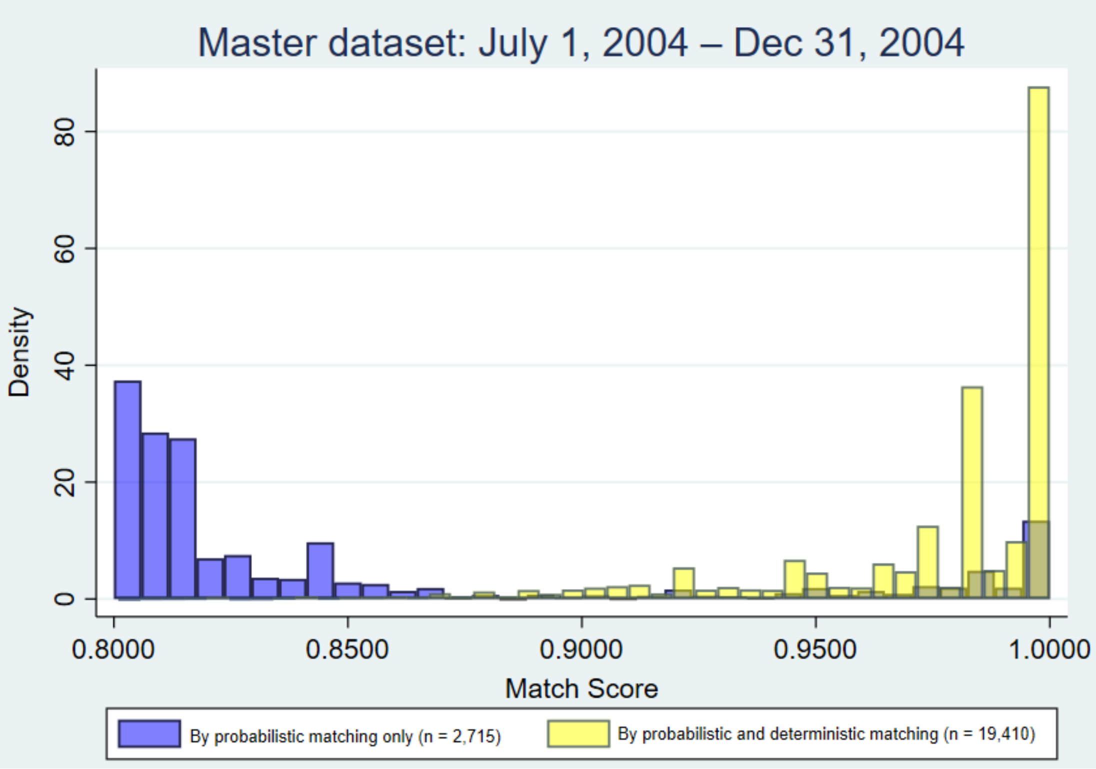
To identify a precise overall match score cut-point for declaring matched pairs as true matches, we deterministically linked each master dataset to its corresponding using dataset using the gold standard. We stratified the matches generated by executing `reclink2` into two groups: (1) matches generated by both probabilistic and deterministic matching, and (2) matches generated only by probabilistic matching. Figure 2 shows the number and distribution of the overall match scores in each of these groups. Except for the last two linkages, all linkages have a similar stratified distribution of overall match scores.

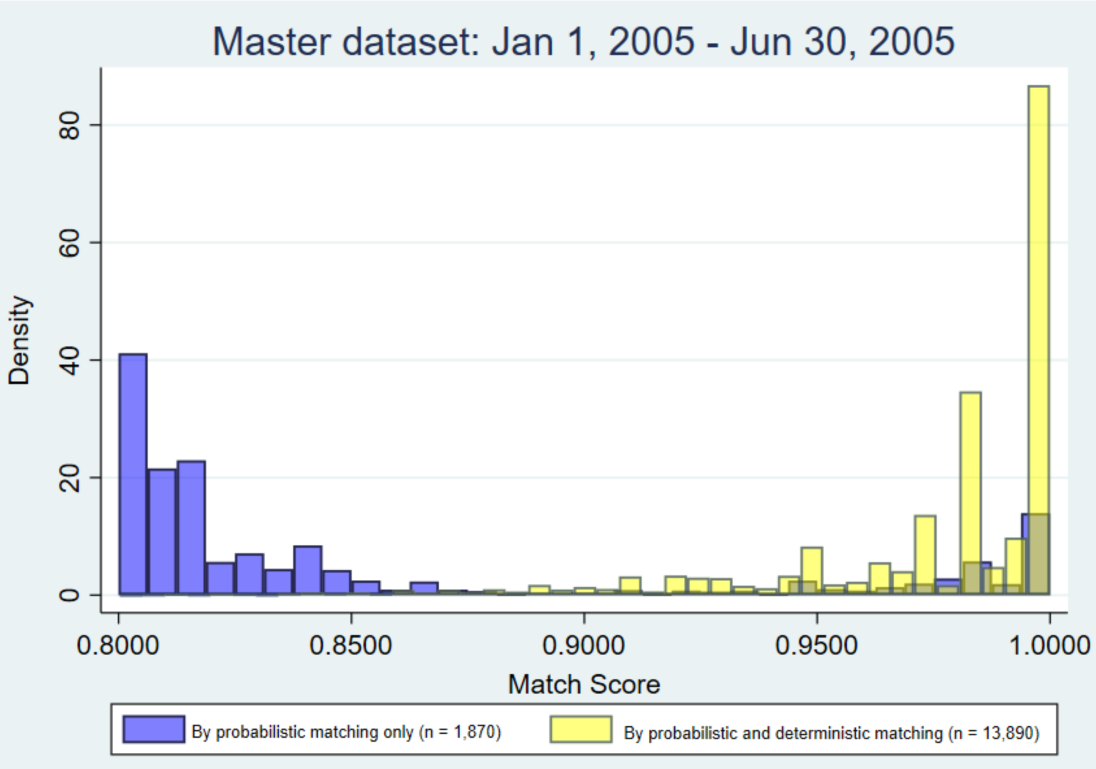
Figure 2 Overlayed histograms showing the stratified distribution of the overall match scores for the matches generated from each probabilistic linkage

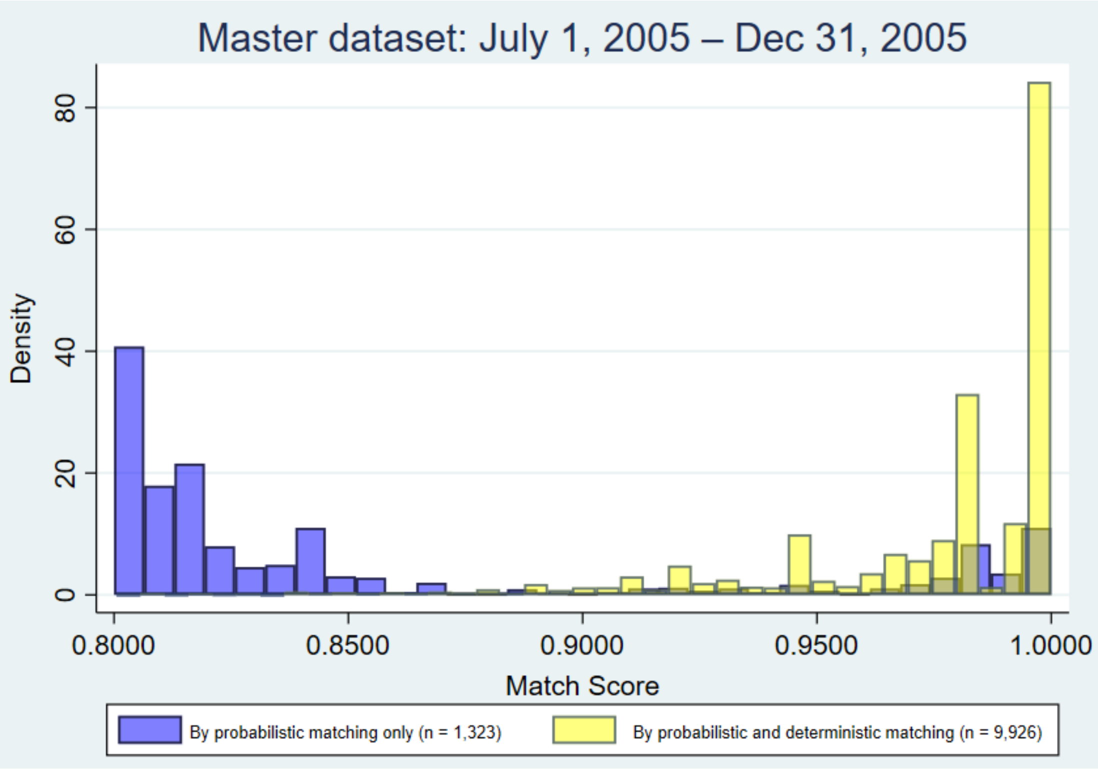


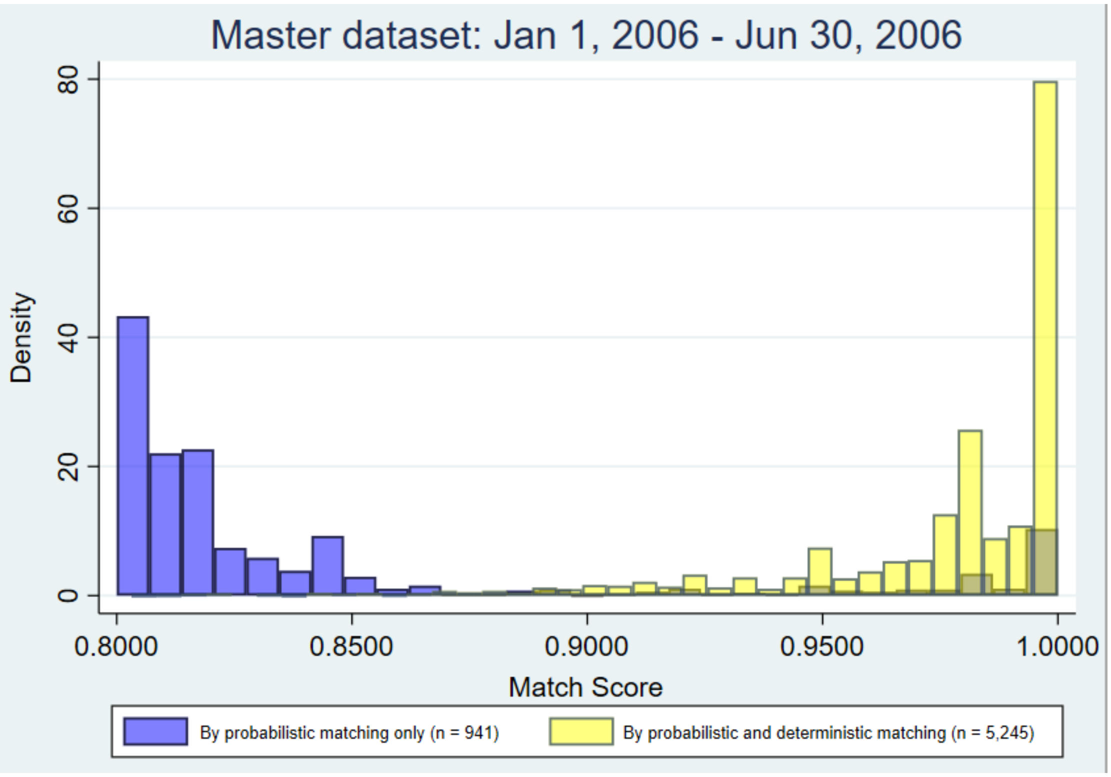

Master dataset: Jul 1, 2003 - Dec 31, 2003
Density
0
20
40
60
80
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
By probabilistic matching only (n = 4,700)
By probabilistic and deterministic matching (n = 29,849)

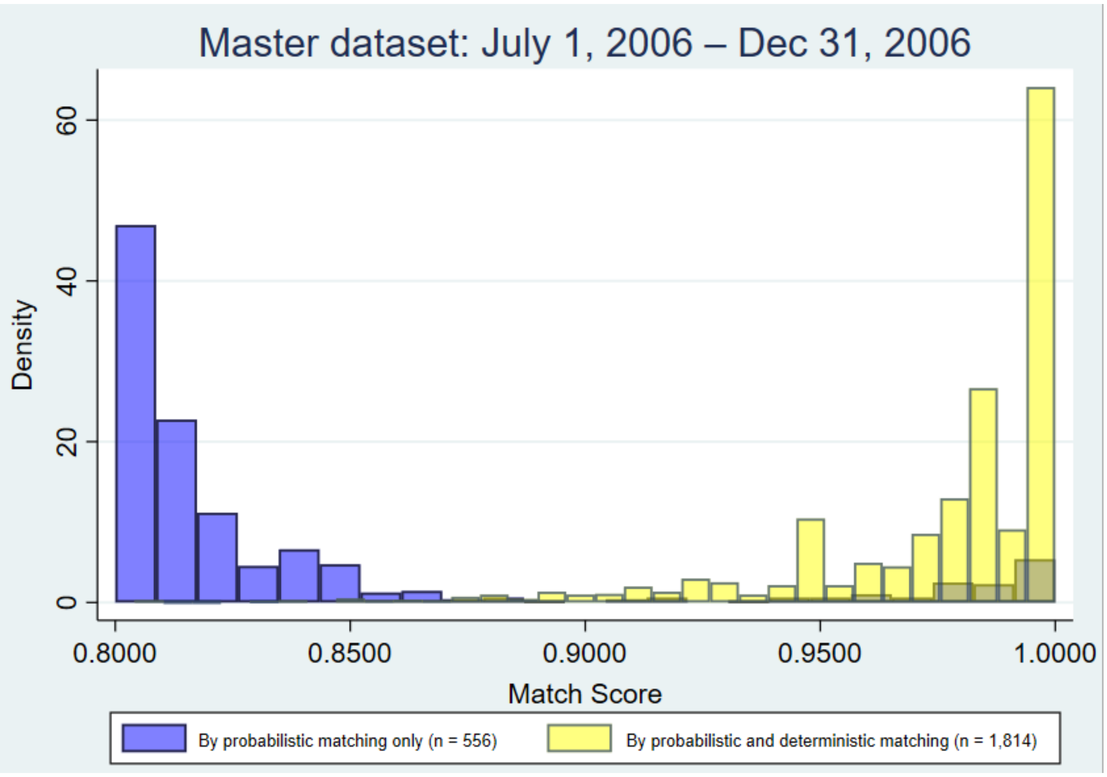

Master dataset: Jan 1, 2004 - Jun 30, 2004
Density
0
20
40
60
80
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
By probabilistic matching only (n = 3,694)
By probabilistic and deterministic matching (n = 23,887)

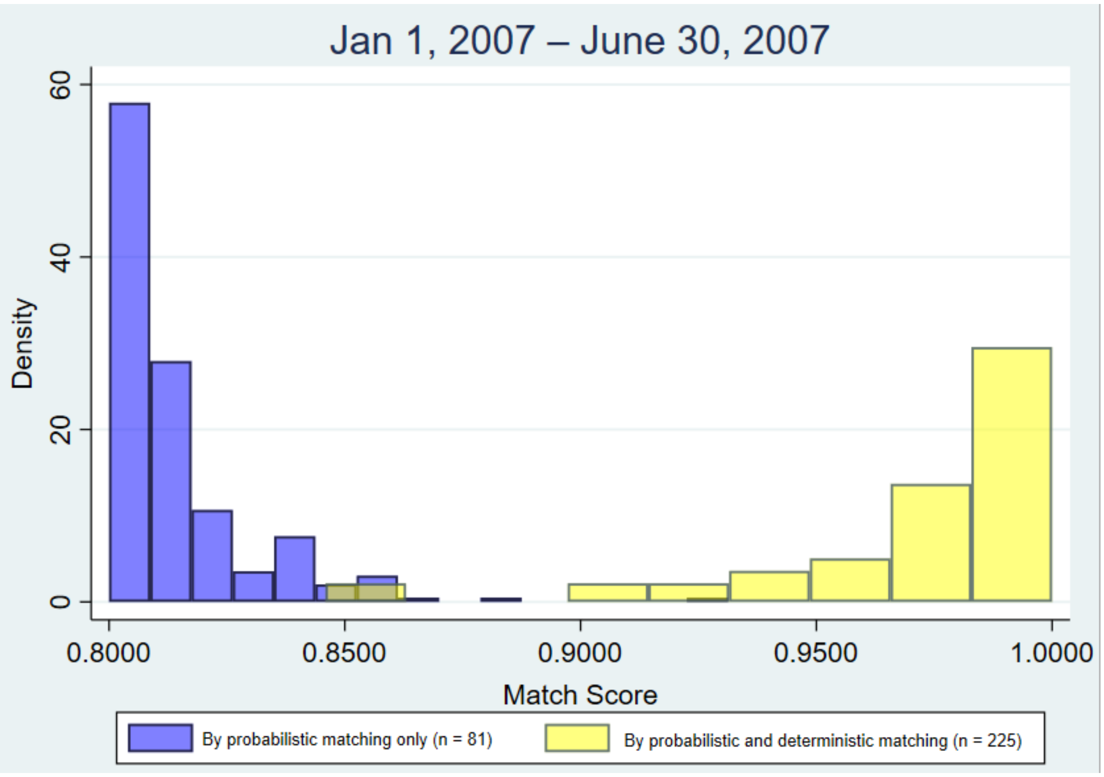

Master dataset: July 1, 2004 – Dec 31, 2004


Master dataset: Jan 1, 2005 - Jun 30, 2005


Master dataset: July 1, 2005 – Dec 31, 2005
Density
0
20
40
60
80
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
By probabilistic matching only (n = 1,323)
By probabilistic and deterministic matching (n = 9,926)


Master dataset: Jan 1, 2006 - Jun 30, 2006
Density
0
20
40
60
80
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
By probabilistic matching only (n = 941)
By probabilistic and deterministic matching (n = 5,245)


Master dataset: July 1, 2006 – Dec 31, 2006
Density
0
20
40
60
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
🟦 By probabilistic matching only (n = 556)
🟨 By probabilistic and deterministic matching (n = 1,814)


Jan 1, 2007 – June 30, 2007
Density
0
20
40
60
0.8000
0.8500
0.9000
0.9500
1.0000
Match Score
🟦 By probabilistic matching only (n = 81)
🟨 By probabilistic and deterministic matching (n = 225)

Matches generated by both deterministic and probabilistic matching are true matches. From Figure 2, we notice that the number of these matches becomes very minimal below a certain match score. Matches generated only by probabilistic matching are either:

1. true matches that the deterministic matching did not generate due to social security number errors. These are the matches at the higher range of match score (*the blue color behind the yellow in the overlayed histograms in Figure 2*).
2. false matches below the appropriate match score cut-point. These could include some true matches as well, but we hypothesize that they account for a very small percentage of those.

For each linkage, we identified an overall match score cut-point that balances between having the highest number of true matches while keeping the false matches at the lowest number possible. All linkages had an overall match score cut-point of 0.87, except the last linkage which had a cut-point of 0.89. We expected all linkage to have a similar overall match score cut-point, as the quality of our data (missingness, errors) is roughly the same over time. Table 4 shows the results of probabilistic and deterministic linkages.

Deterministic linkage generated a total of 139,640 matches. Of these, the probabilistic linkage generated 131,800 matches (94.3%). The number of matches generated by probabilistic linkage but not by deterministic matching was 4,453. These numbers can be illustrated in a 2 by 2 table as in Table 5. Thus, the sensitivity of our probabilistic linkage is 94.3% and the positive predictive value is 96.7%.

Table 4 Results of probabilistically and deterministically matching each master and using datasets

Master dataset		Using dataset		Match score cut-point	Number of matches by probabilistic linkage*	Number of matches by deterministic linkage
Six-month interval	Number of records	Range of years	Number of records			
Jan 1st, 2003 – June 30th, 2003	69,593	July 1st, 2003 – Dec 31st, 2007	642,358	0.87	31,191	32,685
July 1st, 2003 – Dec 31st, 2003	72,585	Jan 1st, 2004 – Dec 31st, 2007	538,867	0.87	29,935	31,038
Jan 1st, 2004 – June 30th, 2004	67,323	July 1st, 2004 – Dec 31st, 2007	441,864	0.87	24,148	24,658
July 1st, 2004 – Dec 31st, 2004	64,130	Jan 1st, 2005 – Dec 31st, 2007	353,781	0.87	19,610	19,866
Jan 1st, 2005 – June 30th, 2005	56,600	July 1st, 2005 – Dec 31st, 2007	277,765	0.87	14,027	14,128
July 1st, 2005 – Dec 31st, 2005	54,726	Jan 1st, 2006 – Dec 31st, 2007	209,137	0.87	10,075	10,017
Jan 1st, 2006 – June 30th, 2006	49,674	July 1st, 2006 – Dec 31st, 2007	149,473	0.87	5,326	5,310
July 1st, 2006 – Dec 31st, 2006	49,792	Jan 1st, 2007 – Dec 31st, 2007	94,408	0.87	1,862	1,849
Jan 1st, 2007 – June 30th, 2007	45,833	July 1st, 2007 – Dec 31st, 2007	46,731	0.89	79	89

*Number of matches with a match score at or above the identified cut-point

Table 5 A classic 2 by 2 table for estimating sensitivity

		Generated by deterministic matching?	
		Yes	No
Generated by probabilistic matching?	Yes	131,800	4,453
	No	7,840	

5.0 Discussion

Conducting research relating to the period between pregnancies remains a challenge due to lack of longitudinal data that follow pregnancy histories over time. Our pilot study demonstrates how a longitudinal maternal dataset can be successfully generated from states' fetal death and birth records. In the absence of a unique personal identifier, we have shown probabilistic record linkage to be a valid method for linking our maternal records.

In this pilot study, we assessed the validity of probabilistic matching that uses a set of nonunique identifiers to link maternal records. We found that our linkage technique has a sensitivity of 94.3% and a positive predictive value of 96.7%. Although we used social security number, the best available unique identifier, as the gold standard to assess the validity of probabilistic linkage, it was not a perfect gold standard as there were likely to be some minor entry errors in its values. Therefore, numbers of false negative and false positive matches are not perfectly accurate. Using an imperfect gold standard is likely to result in underestimation of sensitivity and positive predictive value, thus the sensitivity and positive predictive value of our probabilistic linkage are in fact higher than what we estimated.

Our work can serve as a template for developing similar longitudinal maternal data sources in other states if a unique maternal identifier is not present. We demonstrated how Stata's existing packages, namely `dtalink` and `reclink2`, can be practically used to facilitate probabilistic linkage. We have identified a set of nonunique identifiers with their appropriate match and nonmatch weights. For this weight estimation, we assumed social security numbers to be sufficient to identify the true match status. Given the very large number of records in our dataset, we believe this assumption generated valid weight estimates. Moreover, we have identified appropriate overall match score cut-points for considering linked pairs as matches. Researchers in other states can use an overall match score cut-point of 0.89 when linking the last six-month interval and 0.87 for all other linkages, given they are using the same set of identifiers and options we used for our linkage. We expect similar sensitivity and positive predictive value with state records that have similar levels of missingness and errors.

When choosing the appropriate overall match score cut-point, we aimed to balance between having the highest number of true matches while keeping the false matches at the lowest number possible. Choosing a higher cut-point would result in missing some true matches, while choosing a lower cut-point would result in more false matches. When having a large dataset that generates a large number of matches, we think it is more appropriate to minimize the number of false matches rather than add a few more true matches, so that results from subsequent analyses of these data would be more reliable.

The greatest value in linking states' fetal death and birth records is that within these records are a large number of important variables related to characteristics of the mother and her fetus or newborn, as well as pregnancy and birth outcomes. Thus, the significant implication of our work is being able to conduct research into maternal and child health and answer causal questions requiring longitudinal data. Results from this research can be used to improve health outcomes of mothers and children.

Bibliography

1. Hall ES, Goyal NK, Ammerman RT, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Maternal and child health journal 2014;18:316-25.

2. Kumar RG, Wang Z, Kesinger MR, et al. Probabilistic matching of deidentified data from a trauma registry and a Traumatic Brain Injury Model System Center: a follow-up validation study. American journal of physical medicine & rehabilitation 2018;97:236.

3. Kesinger M, Kumar RG, Ritter AC, Sperry JL, Wagner AK. A probabilistic matching approach to link de-identified data from a trauma registry and a traumatic brain injury model system center. American journal of physical medicine & rehabilitation 2017;96:17.

4. Croft ML, Read AW, De Klerk N, Hansen J, Kurinczuk JJ. Population based ascertainment of twins and their siblings, born in Western Australia 1980 to 1992, through the construction and validation of a maternally linked database of siblings. Twin Research and Human Genetics 2002;5:317-23.

5. Wasi N, Flaaen A. Record linkage using Stata: Preprocessing, linking, and reviewing utilities. The Stata Journal 2015;15:672-97.

6. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. International journal of epidemiology 2016;45:954-64.

7. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets: Linkage of national clinical datasets without patient identifiers using probabilistic methods. Journal of clinical epidemiology 2021.

8. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. International journal of epidemiology 2002;31:1246-52.

9. Kranker K. Faster probabilistic record linking and deduplication methods in Stata for large data files. 2018.