Methods for Identifying Siblings in Administrative Mortality Data

Nicholas Nolte

November 2024

Contents

1 Introduction

Sibling studies are an important tool in demography, as they allow researchers to control for unmeasured factors that siblings share in their household of origin. We have produced two siblings datasets from the Berkeley Unified Numident Mortality Database or BUNMD. The BUNMD contains transcribed data from Social Security applications, claims, and death files for almost 50,000,000 individuals. Both parents' full names are included in this dataset, which we use to identify siblingships.

Two different matching methods were used to make these datasets. The first is called the *Exact Match* method, because it only finds siblings using exact matches on parents' names. The other we refer to as the *Flexible Match* method, and uses the Jaro-Winkler[1] string similarity formula to allow for some typos and spelling mistakes between names. Both of these initial match methods are then passed through the same cleaning process to arrive at the final dataset. The flexible match allows researchers to use a larger sample size of siblings with hopefully minimal losses in accuracy.

2 Description of Datasets

Sibships are identified among individuals in the BUNMD who died at age 65+ in the years 1988-2005. Below we list the final size of each dataset. "Flexible Only" refers to siblingships in Flexible Match that contain individuals not in the Exact Match. The Flexible Match method is almost a strict superset of the Exact Match dataset, as simply allowing some distance between names generally results in finding additional matches. However, some individuals and siblingships exist in the Exact Match but not the Flexible Match, due to the procedure for breaking up larger sibling groups with discordant parent middle names. The Flexible Match adds roughly 1.5 million new individuals, while losing about 57,000 of the 4.8 million in the Exact Match.

Dataset	Individuals	Siblingships
Exact	4,767,193	2,130,398
Flexible	6,252,614	2,745,707
Flexible Only	1,542,349	835,544*

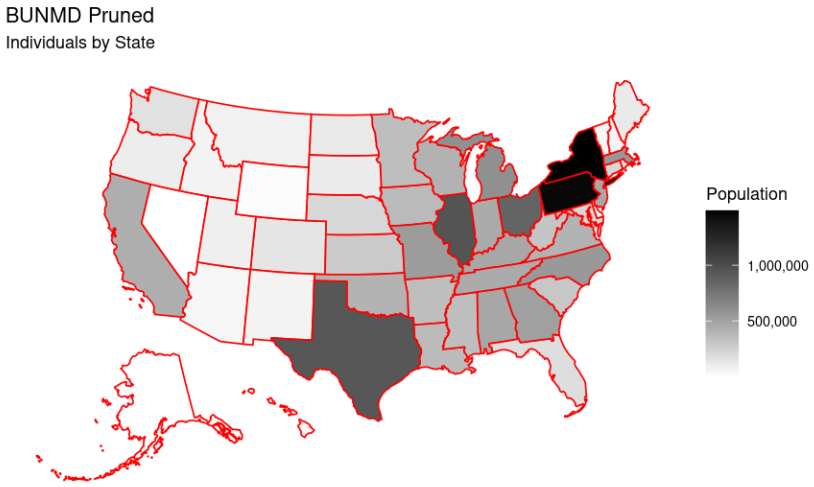
*sibships where at least one sibling is only identified in the flexible match

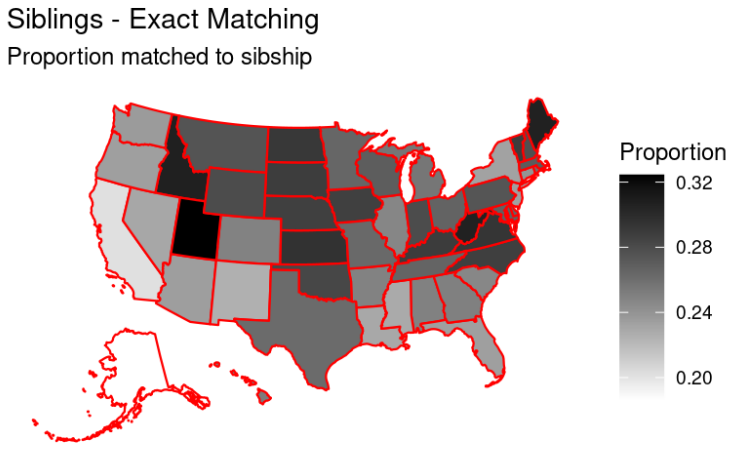
2.1 Geographic Distribution

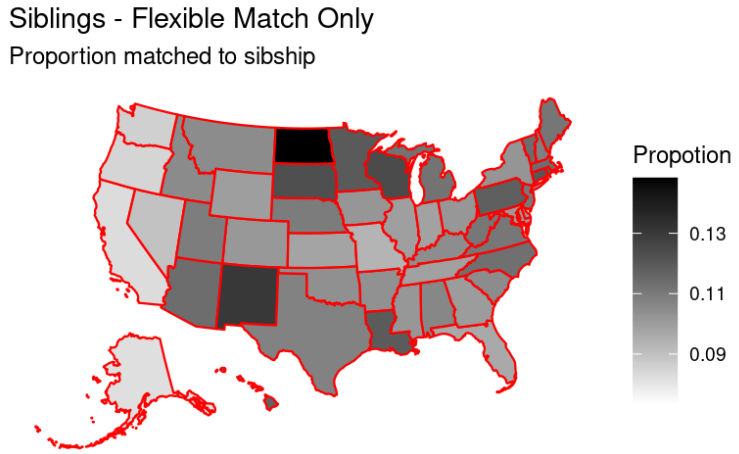
The plots below show the population distribution by birthplace in the BUNMD and the proportion of individuals from each birthplace who are matched to sibships. The sibling data aligns geographically with the relative population sizes in the BUNMD (first map), but the datasets do have different states for

[1] https://en.wikipedia.org/wiki/Jaro-Winkler

which they find proportionally more siblings per population (second and third maps). The most notable states more represented in the Flexible Matches are North Dakota, New Mexico and Louisiana. Researchers interested in these states specifically would most likely benefit more from using the expanded flexible match.

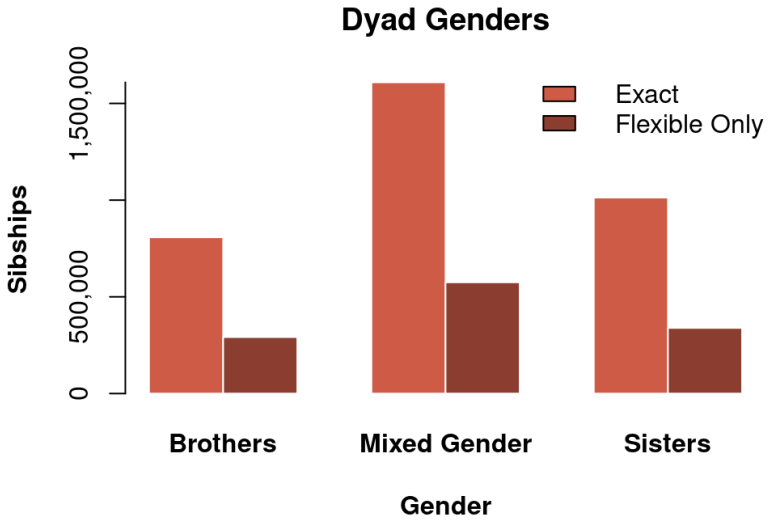






2.2 Dyad Genders

Both datasets have slightly more women due to gendered mortality differentials (women are more likely to survive to age 65), but the genders of sibling dyads are close to the expected probabilities.



2.3 Siblingship Size

Both datasets have a similar distribution of siblingship sizes, with the plot below showing from the Exact Match dataset. The modal sibship size is 2, and the maximum size is 9.

3 Dataset Preparation

Both methods use the same initial string cleaning and standardization steps to prepare the BUNMD for siblingship matching. We use our clean-names.r script to do this.[2] This makes all letters lowercase, removes punctuation, splits middle and first names, and replaces nicknames with full names from a database.

One important point of note with the nicknames process, is that it requires a sex column to properly correct the names. This is not given for the parent columns, but is implied by the terms mother and father. We created a dummy variable with the sex for each to allow the script to run.

3.1 Data Pruning

Before we begin actually matching siblings there are some individuals we remove that would not create good matches. The size of the data set prior to pruning is 49,337,827 entries.

3.1.1 High Coverage Region

We only consider individuals whose deaths occur in the BUNMD's high coverage region. This is the set of people dying at age 65+ between 1988 and 2005 inclusive. All individuals whose age or death year is not in this range are excluded. There were also some individuals whose death fell in this range that

[2] https://github.com/caseybreen/censocdev/blob/master/R/clean-names.R

had unbelievably high death ages, going up to 137. To remove these erroneous data we also remove anyone with a death age above 110.

Of all pruning steps, this step removes the most individuals. In addition to narrowing the sibling search window, this also makes this dataset easier to use for anyone doing research with the BUNMD who would most likely be using the high-coverage mortality data. Following this step the BUNMD has been reduced to 29,422,690 entries.

3.1.2 NA Values

Another large set of individuals we remove are those who do not know some of their parents' first or last names. In addition to entirely blank fields, we remove various NA indicators such as "UNK", "unkn", "not stated" and so on. Further, we remove anyone whose parent's first name is "not". This simplifies the process, as many versions of "Not (Stated, Avail, Listed, Given)" are present and sometimes even misspelled. This leaves 19,874,026 entries remaining.

3.1.3 Single Character Names

It is a similar case when a first or last name is only 1 character long. Single letter names led to many spurious matches, and thus anyone whose parents' first or last name was only 1 character long were removed. This leaves 19,858,035 entries remaining.

3.1.4 Shared Parent Last Name

As we are matching on parents' first and last names, we are matching on 4 distinct variables for each person. One challenging category for this is mother's last name, as her name changed in marriage. While her maiden name should be given, in some cases it seems her maiden name was forgotten or her marriage name was given anyways. Because we are already only matching on four variables, if the mother's name is lost it significantly reduces confidence in matches especially in instances with common names. For example, consider the couples where John Smith marries Anna x. Everyone who forgot their mother's maiden name and left it blank or listed their father's would be lumped together in a clearly incorrect siblingship. Thus we remove anyone who list their parents' last names as being the same. While this may exclude a few genuine cases of marriage with a shared last name, this is fairly uncommon and the benefit outweighs this risk. This solves many problems with sibling formation by removing around 400,000 individuals. Finally, there are 19,408,130 entries remaining that we attempt to match.

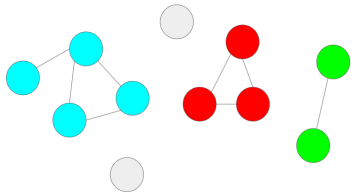
4 Matching Methods

4.1 Exact Match

This is the most restrictive matching method, where siblings are matched only if all of their parents' first and last names match exactly in the BUNMD. This matches 5,264,556 individuals.

4.2 Flexible Match

In this method individuals are blocked into groups based on the first two letters of each parent's first and last names. Anyone with no matches is removed. After this every person is paired with every individual in their group. From here parents' first and last names are given a string similarity score using the Jaro-Winkler formula, and any under a certain threshold are removed. We tried this with different thresholds, and found 0.9 provides the best trade-off between adding new connections while maintaining accuracy. Following this we have a set of linkages between siblings that can be thought of as a graph, with siblings as nodes, and edges between anyone whose parents' names were close enough. Using the R package igraph[3] we take any connected portions of this graph as siblings. This is an efficient way to group the individuals who may not all be linked to every member of their siblingship. Below is an illustration of what this looks like. After this graphical linkage there are 7,132,874 individuals in this dataset.



5 Cleaning

After the initial matching process, we implement some additional cleaning steps to remove less-plausible matches.

The first step in the cleaning process is removing people with multiple social security numbers who have been matched with themselves. These cases are known to exist within the BUNMD, and as they are the same person, they have the same parent's names and end up matched as siblings. We find these people by looking within the groups from the first part and checking either for individuals with same birth and death day or individuals who have the same first and last name. Some twins in theory could have the same birth and death day, but these are exceptionally rare cases that do not affect the size of the overall dataset.

[3] https://r.igraph.org/

Dataset	People Removed
Exact Match	17,106
Flexible Match	22,156

Following this we address issues in the parents' middle names. We only matched on first and last names in the prior stages, so in some cases there are discrepancies in parent's middle name within matched sibling groups. We only look at the first letter of the middle name, as almost all individuals only listed this, and many listed no middle name at all. We do not remove cases where within a siblingship some individuals wrote the same middle name and others wrote none. However sometimes middle initials disagreed within established sibling groups. In some cases these siblingships were broken down further by parents' middle names, and in other cases ambiguous individuals were removed.

Take the following case for example, that looks only at the father's name.

First Name	M. Name	Last Name
John	F	Smith
John	F	Smith
John	[blank]	Smith
John	H	Smith
John	H	Smith

In this case there are clearly two distinct groups of siblings that need to be separated. The individual in the middle, however, is ambiguous, and it is not clear how they should be grouped, therefore they are removed from the dataset at this stage. Also at this stage, siblingships with more than two different distinct middle names listed for parents are removed, as these cases often introduce many ambiguities in sibling groups.

In the next step we remove siblingships that have too large of an age spread among all siblings, as these cases have higher error than siblingships with a smaller age spread. For this we remove groups where the difference between the oldest and youngest member is greater than 15 years. Some siblingships had age gaps as far apart as 30 years, which while possible, is more unlikely.

After this stage we have the final datasets:

Dataset	Individuals	Siblingships
Exact Match	4,767,193	2,130,398
Flexible Match	6,252,614	2,745,707

6 Performance Metrics

We have designed our process in a conservative manner so as to avoid the mistaken assignment of unrelated individuals to the same sibling set.

In addition, we have created a number of metrics meant to demonstrate the accuracy of our sibling dataset. Some of these measures examine the quality of matches via expected correlations among siblings, such as birthplace or race.

Others look at overall dataset bias to determine what states or genders were most common in the dataset and how this compares to initial BUNMD.

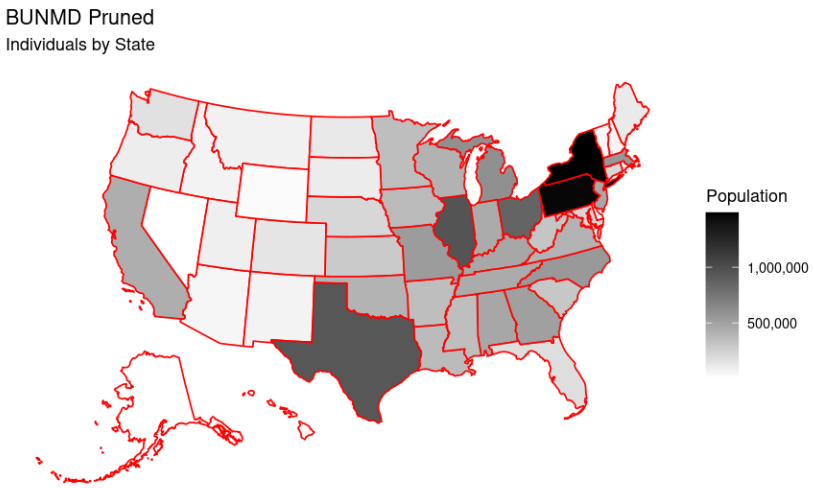
6.1 Gender Distribution

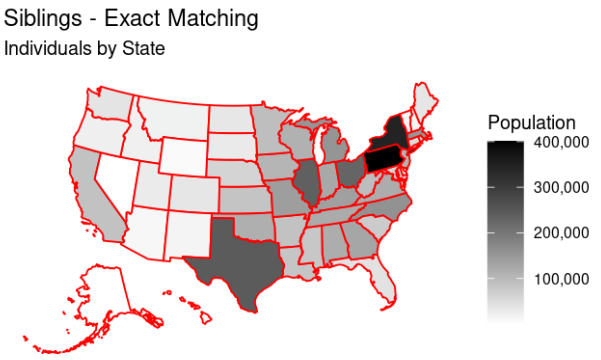
First we look at the fraction male of each of the datasets. We use the pruned BUNMD, which only includes people in the high coverage region, ages 65+ between 1988 and 2005 inclusive. Focusing on older-age deaths leads to the female skewing of this dataset as a whole. Overall the closeness of the dataset to the pruned BUNMD gives us confidence that the datasets have only a small skew in the gender ratio.

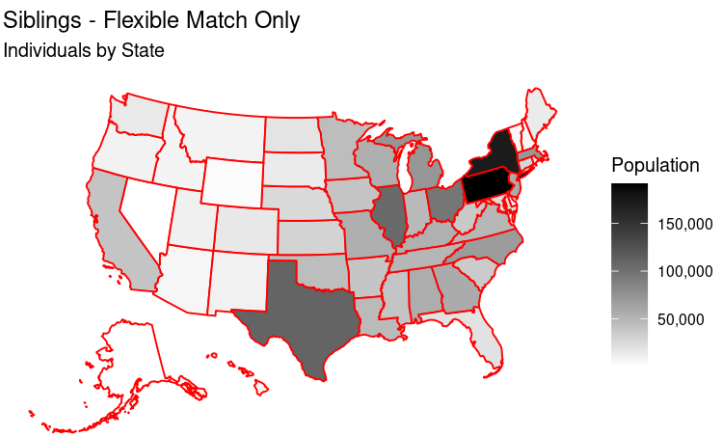
Dataset	Fraction Male
Exact Match	0.4710552
Flexible	0.4750396
Flexible Only	0.4793873
BUNMD Pruned	0.4721718

6.2 Birthplace Distribution

Here we show the birthplace distribution of the siblings we found vs the pruned BUNMD. From these plots we see that the geographic distribution of siblings found generally follows that of the population we selected siblings from.







6.3 Sibling Comparison

Here, we compare matched sibling using non-matching variables that we would generally expect to be consistent within sibships.

6.3.1 Birthplace

While not a guarantee of siblings, we expect fairly high agreement between siblings' birthplaces as families often stay in the same state.

Dataset	Fraction 1 Birthplace
Exact	0.871
Flexible Only	0.844

6.3.2 Race

In the BUNMD we are given the information on the applicants first and last race listed with social security as well as if their race ever changed across their

application. We select those siblings who both have a race listed, and have never changed their race on a filing as the dataset we use to compare siblings.

Dataset	Fraction 1 Race
Exact	0.932
Flexible Only	0.918

6.3.3 Life Expectancy

Because siblings share certain genetic and upbringing similarities, it is likely that they could have more similar life expectancy than individuals chosen at random. In addition to genetic conditions, the similar birthplace and upbringing makes it more likely, but far from guaranteed, that their lifestyles in life may be similar. Per Sebastiani et al. (2015)[4], we see that this is at least the case in old age. Here we explore whether this could be found to some degree at all ages. Thus we expect a certain shared death age for siblings. We show the mean standard deviation of age at death within sibships below.

Dataset	Avg. Standard Dev. of Death Age
Exact	4.666412
Flexible Only	4.830263

The mean standard deviation of death ages among a random pairing of records in the BUNMD was 6.25 years, for comparison.

7 Notes for Researchers

These sibship datasets were created to provide the most accurate sibling identification possible, and not to determine all possible siblings. In addition to the limits of our dataset, this means siblingships here should be expected to be slightly or significantly smaller than the total size. Many true siblingships of 3 or more siblings may only have two siblings marked here or similar reductions. We also make no accounting for cases of half-siblings or adopted siblings. While these cases possibly confound the idea of controlling for genetic factors, they most likely account for a relatively small fraction of the data. In some cases these siblings might be able to be distinguished by a different race, but for individuals of mixed-race this would not be a surefire method, and so it was thought best to leave these cases.

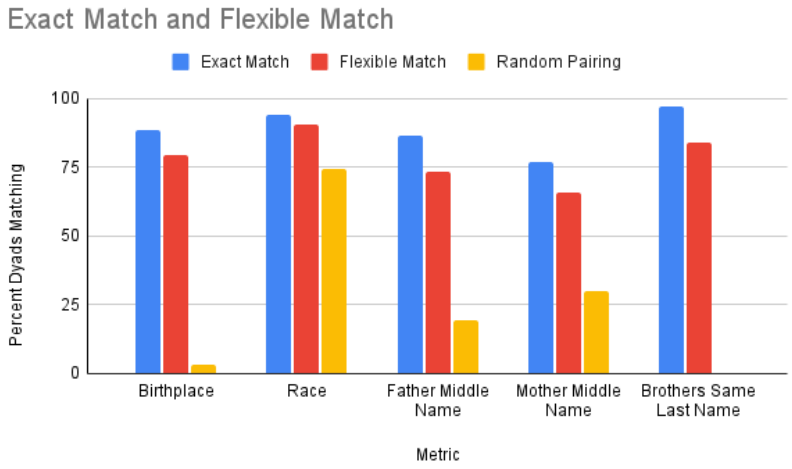
Also note that as this data was made from only the high-coverage portion of the BUNMD, only individuals dying 65+ between 1988 and 2005 are considered. Having this older-age cutoff means that men as well as people of historically disadvantaged racial, ethnic, or socio-economic groups will be less represented. Overall the Flexible Match seems slightly more inclusive of the underrepresented groups by gender and race, but by only a small margin. The

[4]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4757962/

fact that this matching methodology relies on parents' names may also select for more advantaged groups. To be matched to a sibling group, individuals must 1) know, 2) report, and 3) spell all four parent first/last names to a high degree of accuracy, all requirements that may be related to level of literacy.

Taking into account all of the metrics, the best dataset for accuracy is the Exact Match, as it is the most restrictive. While it throws out many good matches over typos, it has better performance on the metrics above. For a larger dataset the Flexible Match does a good job of increasing dataset size without significantly worsening in the metrics.

As a final summary of dataset statistics, we evaluate a number of metrics on our datasets, and compare these to a random pairing from the high coverage region of the BUNMD. Only dyads are included, and the percent given is the percent of all dyads where the given category matches, regardless of whether it is blank.



7.0.1 Possible Improvements

The Flexible Match algorithm was very straightforward, only a simple linear threshold was used, and the same for all parents' first and last names. We found the algorithm to provide a good balance of dataset size and accuracy, finding 32% more siblingships than the exact match method.

The name standardization at the start is also open for improvements. The script lists several additional titles that could be added to help with the list. In addition, a better cleaning could be done of the data, which involves many missing names and various ways of marking this. We do our best to clean these, and none seem to have made it into the final dataset, but a better removal process could still be worthwhile.

This project used only the high coverage region of the BUNMD, and after our pruning there were only around 18 Million individuals out of 50 Million possible people. It would be more difficult to find matches in the subset of

pruned data, as almost all of these cases involved missing parents' first and last names.