BUNMD Sibships User Guide*

Description

These files allow researchers to identify sibships in the Berkeley Unified Numident Mortality Database (BUNMD).

We locate sibships (sibling groups) in the BUNMD by matching individuals on their parents' first and last names as recorded in Social Security Numident records. Two methods are used to to match siblings:

1) The **exact** match method identifies siblings only with exactly identical parent names (after names have undergone cleaning and standardization). This is the most stringent match method.

2) The **flexible** match method permits parents' names to be slightly different, within a threshold defined by Jaro–Winkler string distance, in addition to exact matches. This allows siblings to be matched even in cases of minor misspellings, mistranscriptions, or spelling variations in parents' names among sibships (e.g., mother's maiden name recorded as "Brannum" and "Branum"). This increases the number of siblings found, but has higher potential to falsely match unrelated individuals. Most (but not all) individuals and sibling connections identified using the exact match method are also identified using this method.

Sibships are identified among individuals in the BUNMD who died at age 65+ in the years 1988-2005. They may be of any gender composition and contain from 2 to 9 siblings each. An overview of the size of resulting sibships created by each method is presented below:

Match method	Number of individuals	Number of sibships	Mean sibship size
Exact match	4,767,193	2,130,398	2.24
Flexble match	6,158,636	2,706,858	2.28

For a detailed description of sibling identification methodologies and characteristics of sibships, please see the paper: *Methods for Identifying Siblings in Administrative Mortality Data* available online at https://censoc.berkeley.edu/documentation/.

Usage

Each sibling dataset consists of two columns: a unique individual identifier `ssn` (social security number), and an identifier for each sibling group. The ssn of one sibling within each sibship is used as the group identifier.

```
# Show first rows the exact match
head(bunmd_sibs_exact, 5)
```

```
##        ssn sib_group_id_exact
##      <int>              <int>
## 1: 1010038            1203849
## 2: 1010047            1010122
## 3: 1010077           32055083
## 4: 1010122            1010122
## 5: 1010178            1010185
```

*Last updated: 19 November, 2024

These files must be used in conjunction with the BUNMD. Users will need to merge the datasets using the unique identifier `ssn`, as in the example code below:

```
require(data.table)
# Read full BUNMD
bunmd <- data.table::fread("bunmd_v2.csv")
# Read BUNMD sibships IDs
bunmd_sib_id <- data.table::fread("bunmd_sibs_flexible_match_v1.csv")
# Attach sibling IDs to the BUNMD, keeping all records in the BUNMD
bunmd_with_sibs <- merge(bunmd, bunmd_sibs_id, by = "ssn", all.x = TRUE)
```

We note that there are no person-weights specifically for the subsets of the BUNMD belonging to sibships.