BUNMD Cleaned Names File*

Page	Variable	Label
2	ssn	Social Security Number
3	fname_clean	First Name (Cleaned)
4	mname_clean	Middle Name (Cleaned)
5	lname_clean	Last Name (Cleaned)
6	father_fname_clean	Father's First Name (Cleaned)
7	father_mname_clean	Father's Middle Name (Cleaned)
8	father_lname_clean	Father's Last Name (Cleaned)
9	mother_fname_clean	Mother's First Name (Cleaned)
10	mother_mname_clean	Mother's Middle Name (Cleaned)
11	mother_lname_clean	Mother's Last Name (Cleaned)

Summary: The BUNMD Cleaned Names File (N = 49,337,827) is a supplementary dataset that provides cleaned names for individuals and their parents in the Berkeley Unified Numident Mortality Database (BUNMD). Orignal name data is taken from Social Security Numident application and death records, and can be found in the BUNMD. Cleaning names includes removing non-alphabetic characters, standardizing NA values, and replacing some nicknames and name variants with standardized names. Researchers may attach this file to the full BUNMD using the `ssn` (Social Security number) variable, which uniquely identifies each person in the dataset.

Notes:

1. NA values for names are usually represented with an empty string. However, sometimes strings such as "unknown", "unk", "missing" or "not stated" were present in the name fields of Social Security records, usually for parents' names. Many common words indicated missing data have been removed from this dataset, but some such invalid strings may remain.
2. This process of name cleaning and standardization is primarily designed to facilitate record linkage. Decisions for how to parse names, standardize names, and convert nicknames to full names are largely compliant with the original implementation of the automated ABE linking algorithm developed by Abramitzky, Boustan and Eriksson (2012, 2014, 2017).

*Last updated: 01 March, 2024

ssn

Label: Social Security Number

Description: ssn reports a person's Social Security number, as recorded in Numident death records. It uniquely identifies all records, and can be used to link this file to the complete BUNMD.

fname_clean

Label: First Name (Cleaned)

Description: fname_clean is a character variable reporting the first 16 letters of the person's cleaned first name, as recorded in the Numident death records. This variable was cleaned from raw first names by removing titles (e.g., Dr.) and replacing nicknames with standard names (e.g., Billy to William). Non-alphabetical characters were also removed.

mname_clean

Label: Middle Name (Cleaned)

Description: mname_clean is a character variable reporting the first 16 letters of the person's cleaned middle name, as recorded in the Numident death records. This variable was cleaned by removing non-alphabetical characters.

lname_clean

Label: Last Name (Cleaned)

Description: lname_clean is a character variable reporting the first 21 letters of the person's cleaned last name, as recorded in the Numident death records. This variable was cleaned from raw last names by removing non-alphabetical characters. Multiple-word last names with certain prefixes were combined into one word (e.g., Mc Donald to McDonald).

father_fname_clean

Label: Father's First Name (Cleaned)

Description: father_fname_clean is a character variable reporting the first 16 letters of the person's father's cleaned first name, as recorded in the Numident application records. This variable was cleaned from the father's raw first name by removing titles (e.g., Dr.) and replacing nicknames with standard names (e.g., Billy to William). Non-alphabetical characters were also removed.

father_mname_clean

Label: Father's Middle Name (Cleaned)

Description: father_mname_clean is a character variable reporting the first 16 letters of the person's father's cleaned middle name, as recorded in the Numident application records. This variable was cleaned from the father's raw middle name by removing non-alphabetical characters.

father_lname_clean

Label: Father's Last Name (Cleaned)

Description: father_lname_clean is a character variable reporting the first 21 letters of the person's father's cleaned last name, as recorded in the Numident application records. This variable was cleaned from the father's raw last name by removing non-alphabetical characters. Multiple-word last names with certain prefixes were combined into one word (e.g. Mc Donald to McDonald).

mother_fname_clean

Label: Mother's First Name (Cleaned)

Description: mother_fname_clean is a character variable reporting the first 16 letters of the person's mother's cleaned first name, as recorded in the Numident application records. This variable was cleaned from the mother's raw first names by removing titles (e.g., Dr.) and replacing nicknames with standard names (e.g., Lizzie to Elizabeth). Non-alphabetical characters were also removed.

mother_mname_clean

Label: Mother's Middle Name (Cleaned)

Description: mother_mname_clean is a character variable reporting the first 16 letters of the person's mother's cleaned middle name, as recorded in the Numident application records. This variable was cleaned from mother's raw middle name by removing non-alphabetical characters.

mother_lname

Label: Mother's Last Name (Cleaned)

Description: mother_lname_clean is a character variable reporting the first 21 letters of the person's mother's cleaned last (maiden) name, as recorded in the Numident application records. This variable was cleaned from mother's raw last name by removing non-alphabetical characters. Multiple-word last names with certain prefixes were combined into one word (e.g., Mc Donald to McDonald).