

Supplementary Appendix

Supplementary Appendix

This supplementary appendix presents the procedure for constructing the Berkeley Unified Numident Mortality Database (BUNMD) from the original NARA Numident Records. The BUNMD file contains 28 variables and 49,459,293 death records. See the BUNMD Codebook for variable descriptions, value labels, and tabulations.

NARA Numident Records

In 2013, the Social Security Administration transferred a set of Numident records to the National Archives (NARA). In 2019, we obtained the NARA Numident records, along with their accompanying documentation. The NARA Numident records are a subset of the records in the complete Numident. The NARA Numident records contain three types of entries: applications, claims, and deaths. NARA delivered each set of entries separately as a set 20 fixed-width .txt files ($3 \times 20 = 60$ files in total).

NARA Numident Metadata

We obtained three documents from the National Archives Technical Documentation series: (<https://aad.archives.gov/aad/popup-tech-info.jsp?s=5057>; accessed 11/28/2019):

- Application (SS-5) Records Layout
- Death Records Layout
- Claim Records Layout

The record layout documents contain variable descriptions, value labels, and other technical notes on the variables in the original NARA Numident file. Most importantly, they provide the start and end position for each variable in the fixed-width files.

Code

All data processing was done in the R Statistical Programming Language. The code to construct the BUNMD from the NARA Numident records is available at “[Github.com/caseybreen/wcensoc/](https://github.com/caseybreen/wcensoc/)”.

Pre-Processing

For each of the three entry types, we read in the 20 fixed-width .txt files using the column position specified in the record layout documents. We then appended the 20 files into a single file, creating a single file for each of the three entry types.

We took the following steps to clean each file:

1. We changed the variable names to be more concise and informative. For example, we renamed “NUMI_RACE” variable to “race.”
2. We harmonized the different codes to represent a missing value (“Unknown”, “Unk”, “Un”, and “0”) to “NA.”

Combining NARA Numident Records into the BUNMD

The goal in constructing the BUNMD was to combine the NARA Numident records into a single, harmonized file with one record per person. The original records contain over 100+ variables, many of which are not of general interest to the research community. Further, many of these variables are either completely empty or contain 99%+ missing values, as displayed in Figures 2-4. We select a set of general-interest variables with high completeness.

While a person can only have one death entry, they might have several application or claims entries. Some information may be recorded several times; for example, sex is reported in the application, claim, and death entries. Occasionally, a person reports different values of sex, race, place of birth, etc. across entries. To handle this response inconsistency, we developed a set of decision rules to select the a single value across entries (see Table 2). We denote this process “condensing” — selecting a single best value across entries. In order to study name changes, race changes, and other features, the original NARA Numident records are useful, and are available upon request.

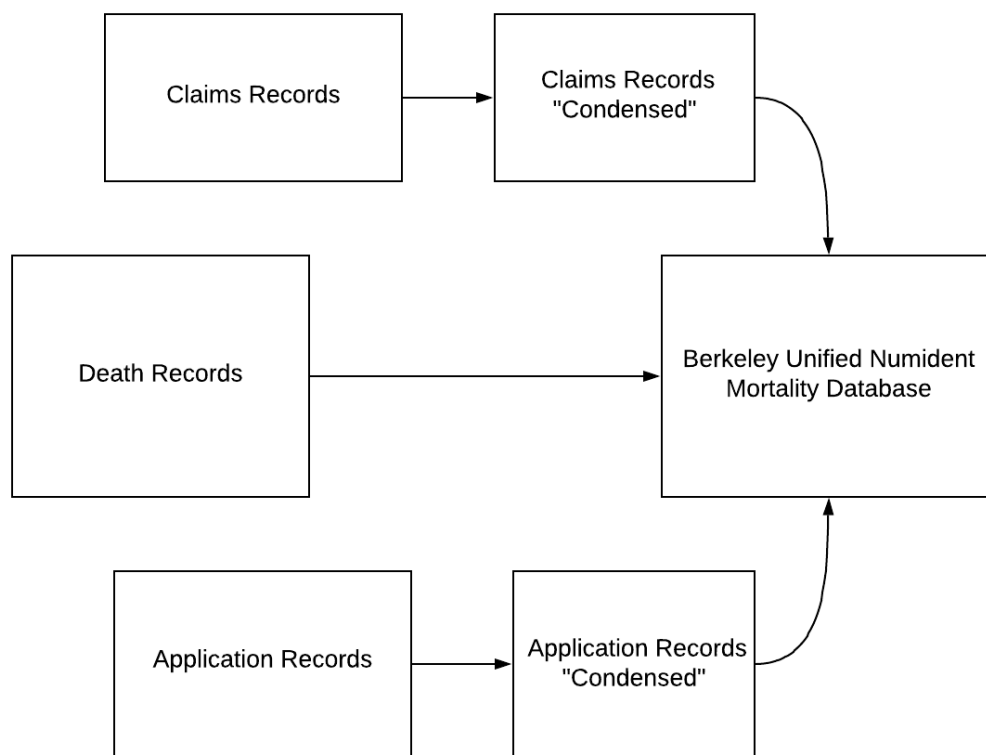
Geography

Place of Birth: There are several geography variables in the NARA Numident records. The application entries have information on birthplace. For persons born in the United States, the geographic resolution is state-level. For persons born outside of the United States, the geographic resolution is country-level. The NARA Numident uses two variables to convey birthplace. The first variable denotes whether a person was foreign born. The second variable contains a two-letter state or country abbreviation. We harmonize these two variables into one variable with a numeric coding schema. This coding schema matches the IPUMS-USA BPLD (Birthplace, detailed) schema.

Place of Death: The Numident death entries contains the 9-digit ZIP Code of the residence at the time of death. Sometimes, the full 9-digit ZIP Code is not available, and an “x” is used to represent a missing digit. This is the original convention used by the Social Security Administration.

Social Security State: The first three digits of a Social Security number correspond to the state in which a Social Security number was issued (prior to 1973) or to the ZIP Code of the mailing address listed in the Social Security application (after 1973). We constructed a variable “socstate” that reports the the State corresponding to the first three digits of the Social Security number. The SSA change the assignment process in 2011 — after the last Social Security number for a person in the BUNMD was issued — and the first three digits no longer correspond to a state.

(<https://www.ssa.gov/employer/stateweb.htm>; accessed 11/28/2019)



(a)

Figure 1: BUNMD Construction Flow Chart

BUNMD Samples and Weights

We create two BUNMD samples with high death coverage. Sample 1 includes deaths to persons age 65+, occurring between 1988 to 2005, from the birth cohorts of 1900 to 1940. Sample 2 is the subset of Sample 1 records with complete information on sex, birthplace, and race. For each sample, we constructed inverse probability weights to the Human Mortality database on age at death, year of birth, year of death, and sex. We broke the sample into cells cross-classified by year of birth, year of death, age at death, and sex. We weighted each cell to the HMD “Deaths by Lexis triangles” totals. This allows aggregation to HMD by period and cohort.

$$W_j = \frac{\text{HMD deaths in cell } j}{\text{Numident Sample 1 deaths in cell } j} \quad (1)$$

Decision Rules used for BUNMD

Variable	Numident Source	Decision Rule
ssn	Death Entry	-
fname	Death Entry	-
mname	Death Entry	-
lname	Death Entry	-
byear	Death Entry	-
bmonth	Death Entry	-
bday	Death Entry	-
dyear	Death Entry	-
dmonth	Death Entry	-
dday	Death Entry	-
zip_residence	Death Entry	-
sex	Death, Application, or Claim Entry	Most Recent
race	Application Entry	Most Recent
bpl	Application or Claim Entry	Most Recent
father_fname	Application or Claim Entry	Maximum Characters
father_mname	Application or Claim Entry	Maximum Characters
father_lname	Application or Claim Entry	Maximum Characters
mother_fname	Application or Claim Entry	Maximum Characters
mother_mname	Application or Claim Entry	Maximum Characters
mother_lname	Application or Claim Entry	Maximum Characters
race_change	Constructed	-
death_age	Constructed	-
socstate	Constructed	-
age_first_app	Constructed	-
number_apps	Constructed	-
number_claims	Constructed	-
weight	Constructed	-
cweight	Constructed	-

Table 1: The decision rules used to construct the BUNMD. For a given variable, we selected values from the Death record if available. If it wasn't available in the death record, we used a decision rule to select it from the Application record. If it wasn't available in either the Death or Application entry, we selected it from the Claim record.

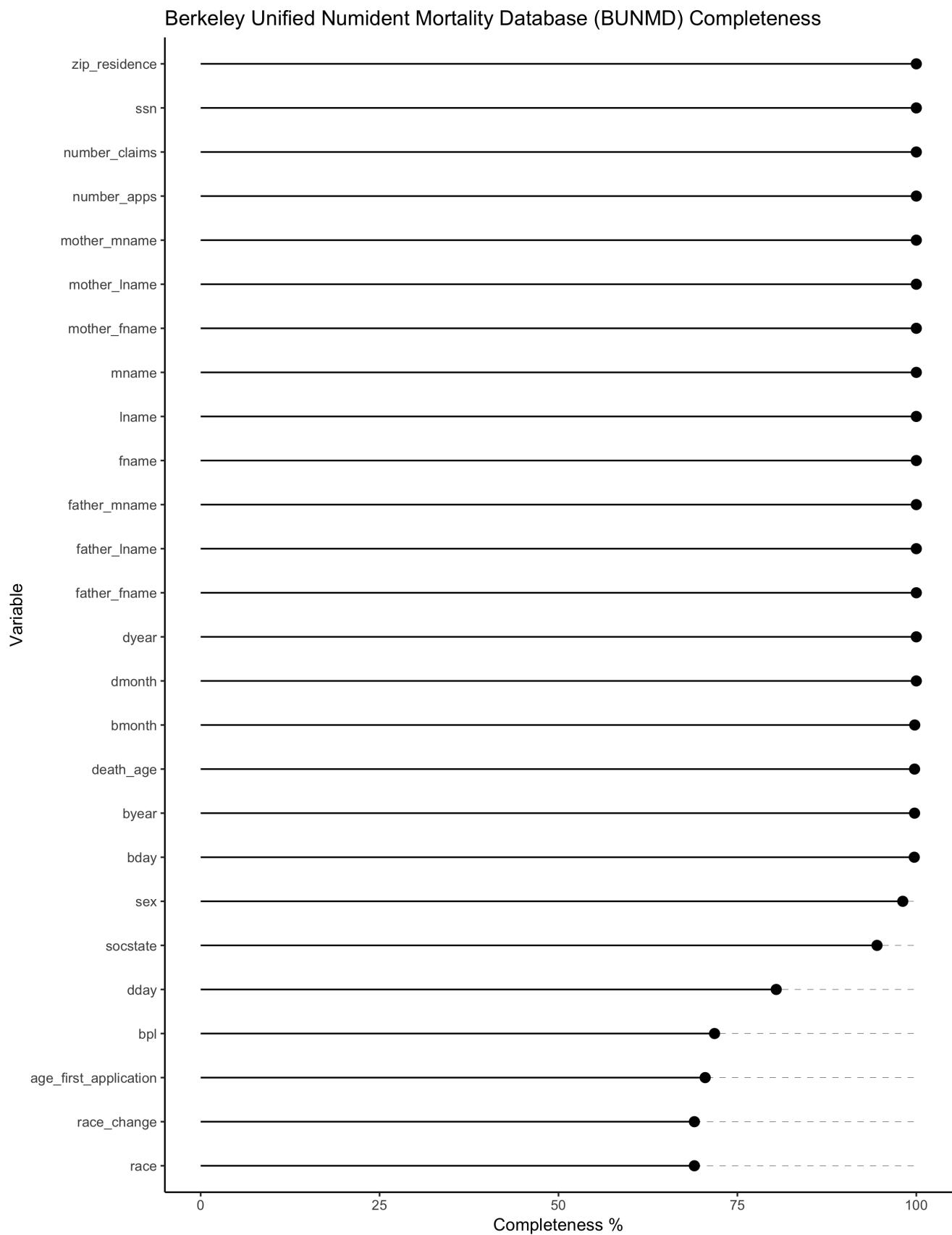


Figure 2

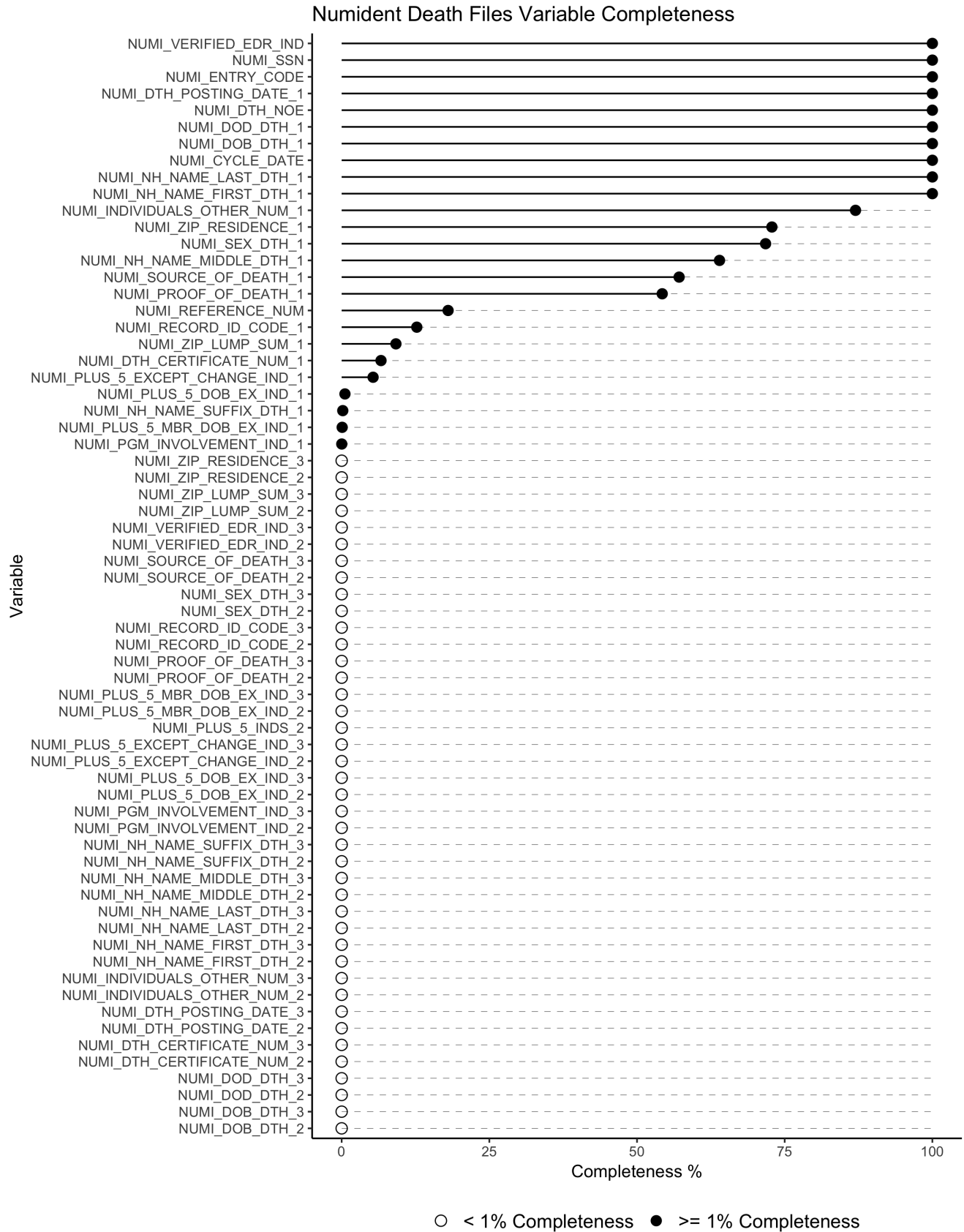


Figure 3

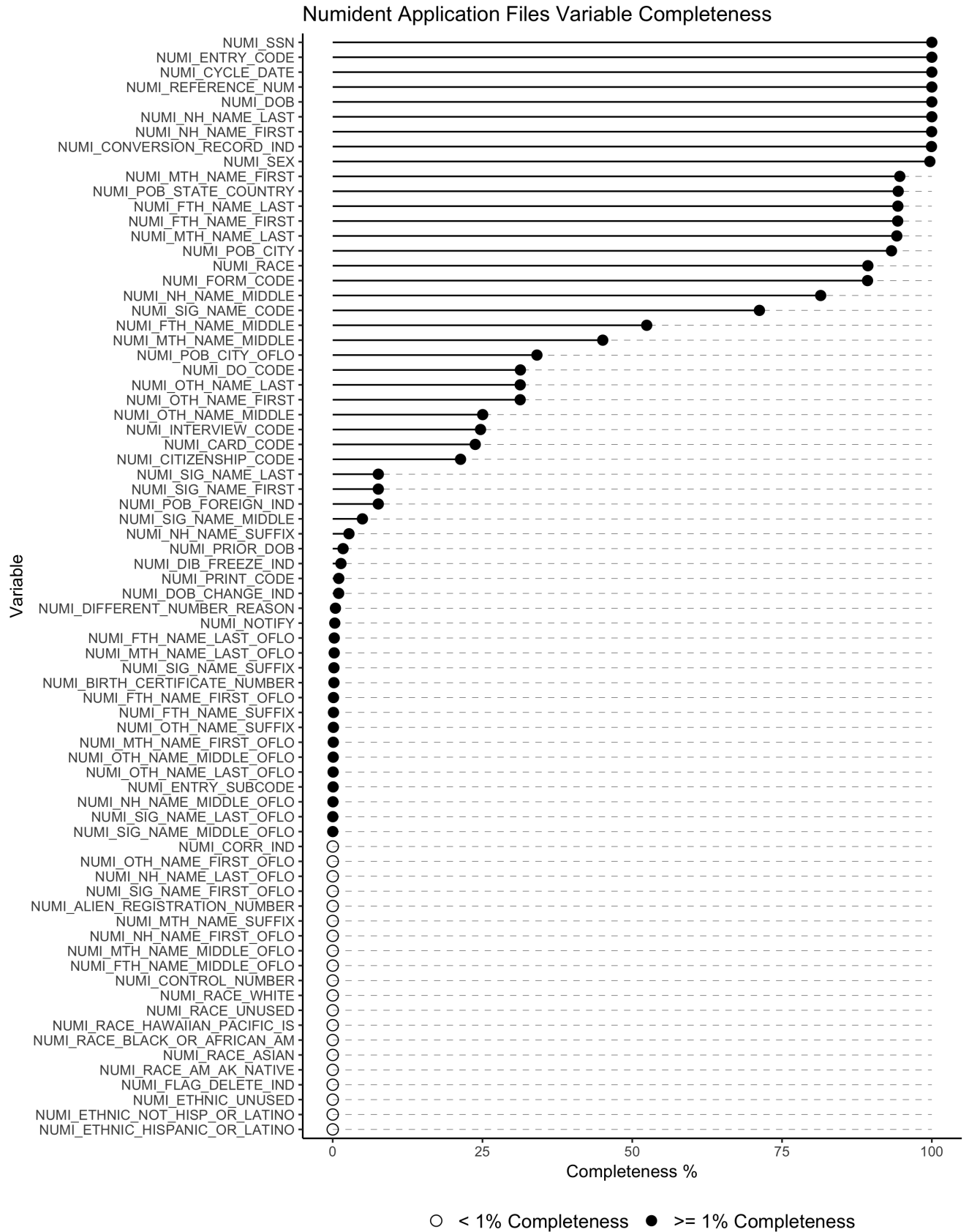


Figure 4

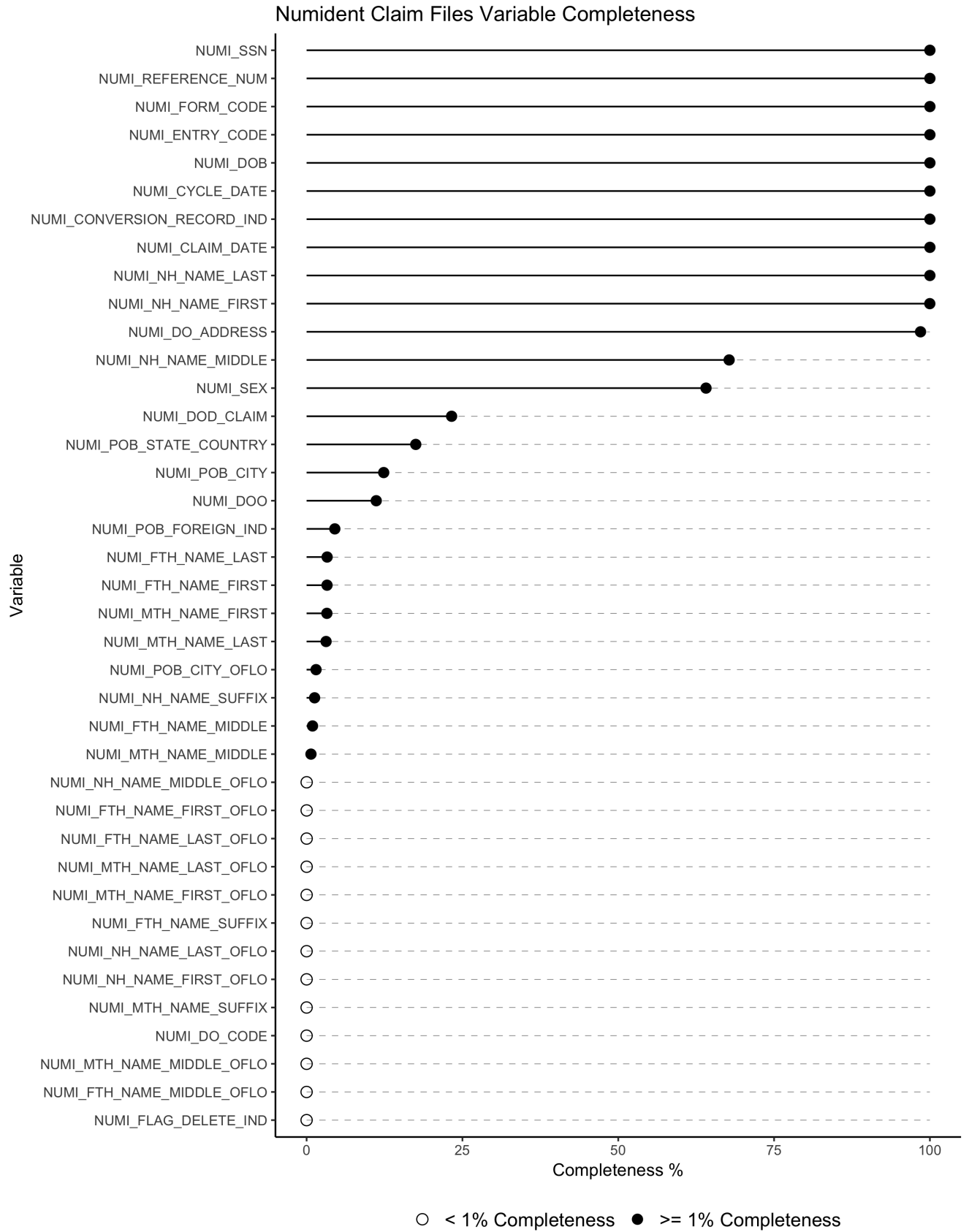


Figure 5