Panel Study of Income Dynamics:

1968-2021 Mortality File Documentation

Release 1

Survey Research Center

Institute for Social Research
The University of Michigan
Ann Arbor, Michigan

July 2023

The 1968-2021 Panel Study of Income Dynamics (PSID) Mortality File and Documentation were created with funding from the National Institute on Aging (R01 AG040213). We also gratefully acknowledge funding for the PSID from the National Science Foundation and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. This document was prepared by Vicki A. Freedman, Robert F. Schoeni, and Kimberly Schlegel, and updated by Esther M. Friedman, Noura Insolera, and Flannery Campbell in July 2023. Suggested Citation: Panel Study of Income Dynamics: 1968-2021 Death File Documentation. Release 1. Institute for Social Research, University of Michigan, July 2023.

TABLE OF CONTENTS

Section I: Overview

1968-2021 Mortality File

The 1968-2021 Mortality File is a restricted file containing information about the deaths of individuals in the PSID who are known to have died from the beginning of the study through the 2021 wave of data collection.

Records on this file include individual identifiers (linkable to the 1968-2021 Individual File), the year in which the death was discovered, date of death, age at death, the state or country in which the individual was born and died, and an accuracy indicator for date of death. For deaths that occurred between 1979 and 2021 additional variables describe matches to the National Center for Health Statistics' (NCHS) National Death Index (NDI) and provide cause of death for cases deemed to be a good match.

Because individuals' death dates are potentially identifying for their survivors, the 1968-2021 Mortality File is only available through the PSID under restricted data contract. For information about obtaining restricted data through secured contact, please see the online documentation or email psidhelp@umich.edu.

Individuals for Whom the Data are Available

Information is available for individuals whose deaths have been identified during the course of the study, that is, for all individuals who are coded as deceased (values 2 and 3) on variable ER32004 (Whether Moved out or Died) on the 1968-2021 cross-year Individual File. The main source of information on whether an individual has died is the yearly nonresponse indicator. In addition, the PSID's major recontact efforts in 1993 and 1994 as well as an Attritor Tracking project undertaken in 2007 determined mortality status for additional persons. For a detailed description of the Attrition Project see http://psidonline.isr.umich.edu/Publications/Papers/tsp/2008-03_PSID_Sample_Leaver_Tracking.pdf

Background on the Development of the PSID Mortality File

The concept of a PSID Mortality File originated from the recognition that as a nationally representative panel study of all ages, PSID could be a valuable source of information on health and mortality in the United States (U.S.). The original file was an outgrowth of a data collection and analysis project during the mid-to-late 1980s funded by the National Institute on Aging and the National Science Foundation. This project obtained date and location of death from the National Death Index (NDI) for individuals in the PSID who were known to have died through the 1984 wave. The deaths of PSID individuals were initially identified either through the yearly 'Reason for Nonresponse' variables or from other sources, such as returned postal material or information provided by surviving family members. This initial NDI match had a 75% success rate.

As part of this initial project, when a satisfactory match to the NDI was found, a subsequent search at the county level for copies of death certificates was attempted. If the search was successful, the certificates were sent elsewhere for coding cause of death. The PSID never received the certificates or coded information, and this information was destroyed at the end of the project, as originally agreed upon as a condition of the contract.

However, the death dates that were collected were preserved and used to create an initial PSID Mortality File. All known deaths through the 1984 wave and subsequently discovered deaths were compiled into one file. Beginning in 2005, PSID began to regularly link all known decedents since 1979 to the NDI to obtain cause of death.

Section II: Structure of the File

Number of Records and Sort Order

The 1968-2021 Death File contains a total of 7,813 records, one for each known deceased individual. The file is sorted, in ascending order, by "1968 Interview Number" (D002) and "Person Number" (D003).

Variables on the 1968-2021 Mortality File

The 1968-2021 Mortality File contains a total of 104 variables, which fall into one of five groups. 1) identifiers and death information from PSID (D001-D014); 2) information sent to NDI as part of the matching process (D015-D024); 3) information received back from NDI as part of the matching process (D025-D051); 4) variables created as part of the in-house determination of confirmed final match (D052-D058); and 5) coded cause of death information received from NDI for confirmed final matches (D059-D104). The following description provides an overview of each group of variables as well as some description of the matching process. Note that variables may have been renamed from previous versions of the release files; see the latest codebook for detailed information for each variable.

Unlike other PSID data releases, this data file contains character data fields as well as numeric data fields; this was deemed necessary as the coded NDI cause of death data contain fields which could not be recoded into numeric equivalents.

1) Identifiers and Death Information from PSID (D001-D014)

In addition to individual identifiers (D002 and D003), date of death is provided from several sources. The first set of mortality data, D006/D007, is coded from PSID sources. These variables incorporate non-response variables, notations from coversheets, returned postal items, obituaries sent by interviewers, and volunteered bits of information from surviving related family members. The second set, D008/D009/D010, was retrieved from the death certificate matching project from the 1980s, described above. The third set, D011/D012, contains a "best" death date, selected by PSID staff as the more accurate from either D006/D007 or D008/D009. (The date of death information received from NDI, described in the next subsection, is not used to determine "best" death date.)

If the exact year of death is unknown, the “best” year variable (D012) contains a range of years in which the death could have occurred. The first two digits represent the earliest possible year, and the last two digits represent the latest year possible. For example, a value of 7294 indicates that death could have occurred as early as 1972, but no later than 1994. These values were based on when an individual last responded to the survey and when the death was discovered.

Two location variables, state in which born (D005) and state in which died (D014), are also included. The birth and death locations were collected beginning in 1996; deaths discovered prior to that wave contain missing data (code 99).

2) Respondent information sent to NCHS (D015-D024)

In order to make a match to NDI, NCHS requires submitted records to include one of the following combinations of information: 1) first and last names and Social Security numbers, 2) first and last names and birth months and years, or 3) social security number and date of birth and sex. We provide NCHS with all available combinations of information meeting these criteria. For deaths through 2011, we used the online Social Security Death Index (SSDI), which draws upon the Death Master File (DMF), to look up Social Security Numbers for known decedents. The DMF lookups have not yet been implemented for deaths discovered after 2011 because of changes in access to and coverage of the DMF. (See section 4 below for a figure showing how the quality of the match has changed over time).

D015 and D016 are indicator variables that describe the submission status and clone status of each individual PSID record. A case was cloned (one record submitted for each unique name) if there was a hyphenated name or if there were multiple last names for a PSID individual. A case was also cloned (one record for each possible month) if month of birth was missing in the PSID data. D017-D024 are variables submitted to NCHS as part of the matching process (formatted according to NDI standards). The parenthetical reference in the variable label refers to each field's location in the NDI user file as it was originally submitted; for a detailed description please refer to http://www.cdc.gov/nchs/ndi.htm .

As of Release 1, the records newly discovered as of 2021 processing have not been submitted to NDI. As per our protocols, we will plan a 2nd Release of the Death File once data are returned from NDI for these new records.

All name fields, the day of birth field, and the Social Security number field, which were submitted to NDI as part of the matching process, have been removed.

3) Matching information received from NCHS (D025-D051)

Variables D025-D051 contain information received from the NCHS for the confirmed matched record's from the NDI selected by PSID. Note that discrepancies may exist between variables obtained by PSID and the corresponding matched NDI variables; no attempt was made to reconcile conflicting information.

All of these fields are coded exactly as NDI provided them, with the following exceptions: D030 (year of death) is converted from a 2-digit to a 4-digit field; D029-D030 and D049-D051 are recoded as numeric fields, with missing data differentiated for cases not submitted to NDI (coded as zero) versus those for which a confirmed match was not selected or not provided by NDI (coded as nines).

Variables D031-D045 indicate whether the PSID record and the selected NDI record match for the given field. However, values of these fields are not provided. For example, D036 indicates whether or not each digit of the SSN matches between the PSID record and the corresponding NDI matching record, but does not provide the actual value of SSN. There are 15 fields altogether.

D046-D048 provide information on the match: whether all items submitted match exactly; the sequence number of the record selected out of up to 50 possible matches; and the number of possible matches provided by NDI to PSID for a case.

D049 is the weighted probability score created by NDI to indicate the quality of the match between the PSID record and the corresponding NDI record. D050 and D051 (also provided by NDI) further summarize this score into categories. For additional background on how NDI creates and utilizes the probabilistic score (D049), refer Appendix A of the NDI User's Guide.

4) Variables created as part of the in-house matching process (D052-D058)

NDI provides up to 50 possible matches for each user record submitted (D048 and D055 summarize this information, with "cloned" cases with multiple records submitted by PSID having up to 600 possible matches). PSID then determines which, if any, potential match is deemed to be the single relatively-best match and the quality of that match.

In making this determination, PSID relies upon the weighted probabilistic score (D049), summary indicators (D050-D051), whether date of death matches (D052), and the sequential ranked order of records provided by NDI (D047). A hierarchical sort is performed so that the single relatively-best match is selected.

The determination is made in several steps. First, possible matches are sorted into subgroups based upon the summary indicator provided by NDI (D051) and a secondary inclusion criteria developed by the

Health and Retirement Survey (HRS). The HRS secondary inclusion indicator is set to 1 if all of the following conditions are met: Matching sequence (D047) >=3 (i.e., one of the top three potential matches); year of birth (D039) is an exact match or is within plus or minus 1 year; and probabilistic score (D049) must be 21 or higher.

Next, a single relatively-best match is selected among the set of "relatively-best" matches. Many cases have just one relatively-best match, which we decide to accept or not (see below) as the match. Other cases (mainly cloned cases and cases with common names) have more than one relatively-best match that are equivalent on several of the factors of interest (see D056). For these cases, we decide which one is the single relatively-best match based on a second hierarchical sorting process that differentiates among relatively-best records with similar characteristics. In order, the factors considered are: probabilistic score (D049), matching sequence (D047), 15-item unweighted sum of matching items (D054), date of death match (D052), and death certificate ID number (an unreleased variable that allowed us to determine if duplicate death records were being considered). The record determined to have a better value is selected as the single relatively-best match for final release. We assign to D057 the variable breaking such "ties" to identify the single relatively-best match. Most cases are resolved by differences in probabilistic score. However, a few cases fail to resolve after all factors are considered. For these unresolved cases, we do not identify or release a single relatively-best match.

To illustrate how the hierarchical matching works, we provide the following example. Suppose for a particular PSID case, two possible matching records in a subgroup (i.e., records having the same value of 1 for statuscode and of 1 for the HRS secondary inclusion criteria) had the same value for probabilistic score. We would then check the value for matching sequence. If the two cases had the same value for matching sequence, we would check the 15-item unweighted sum of matching items (and so on until we identify which of the two matches is better). Once we encounter a variable that indicates which match is better, we assign D057 the variable that breaks the tie.

Finally, variable D058 summarizes the final status for each PSID record submitted to NDI. Analysts can use this information indicating the quality of the match to determine whether to use cause of death information for any given PSID record. The following categories are provided:

> "Best" matches have a single uniquely identified best match with statuscode=1.
>
> "Good" matches are those cases with more than one relatively-best match with statuscode=1. We select a single best match from among these possibilities based on additional information. Because it is possible that we did not select the correct match, we classify these cases as a good match rather than best match.
>
> "Fair" matches were selected from instances where more than one relatively-best match was identified all with statuscode=0.
>
> "Unresolved" matches are those described in the matching process (see previous paragraph) where multiple relatively-best matches existed and a single best match could not be identified.
>
> "Poor" matches are those cases whose relatively-best match(es) failed to meet the secondary inclusion criterion identified by HRS.
>
> "No matches" are those cases that were submitted to NDI but had no possible matches returned.

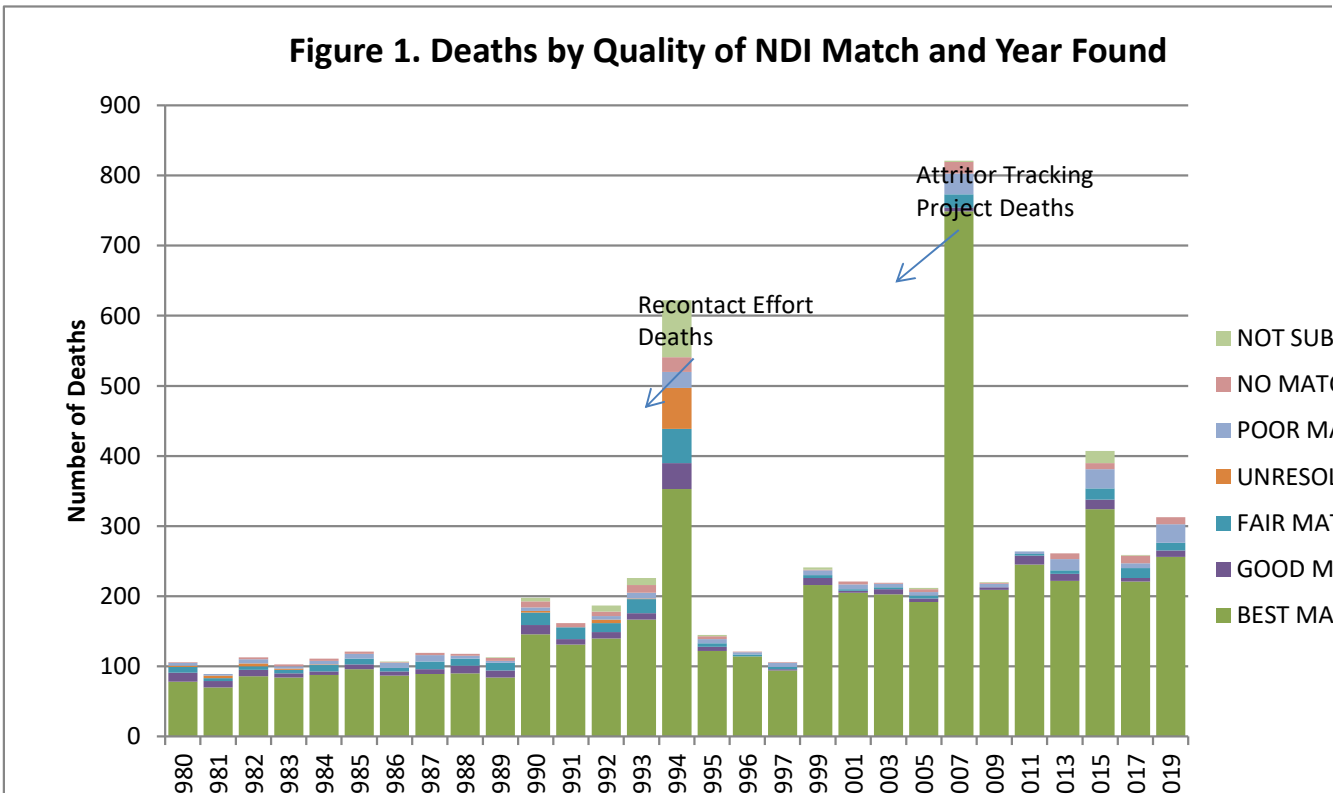
The remaining cases were not submitted to NDI for matching because NDI recordkeeping begins in 1979, and their year of death was 1978 or earlier or because not enough information was obtained to send to NDI.

We validated this summary variable by comparing values for D012 (PSID "best" year of death) and D030 (NDI year of death) by each category of D058 for records with non-missing, valid values on both D012

and D030. Among the "best" matches (D058=1), more than 97% had the same value for D012 and D030; 74% of the "good" matches did so, and just over half of the "fair" matches did so.

Across all deaths that have been found since 1980, 82% were classified as having a "best" match, 4% as "good" and 5% as "fair." A little more than 1% were unresolved, 4% were "poor," 2% had no match, and 2% did not have enough information to be submitted.

Figure 1 shows the distribution of match status for deaths found since 1980 by year found. Substantial numbers of deaths were uncovered in 1994 during a large recontact effort and also in 2007 at the conclusion of the Attritor Tracking Project. Also the number of deaths is higher for the post-1997 period, coinciding with PSID's change to data collection every two years.



Note that the loss after 2011 of access to the DMF file to obtain Social Security Number for decedents led to only a small drop in the percentage with "best" match (from 92.8% in 2011 to 85.4% in 2013). We will continue to monitor quality and may reinstate searches for SSN for cases with less than a "best" match after the DMF data become available (currently after a 3-year delay).

5) Coded cause of death information (D059-D104)

For each match that was determined to be a fair-or-better match (D058, see discussion above), coded cause of death information is included from NDI for those cases where it is available. Note that for a few fair-or-better matches this information was not available because NDI provides cause of death information only for a subset of the possible matches. Refer to the following documents (included on the CD) for a complete description of the codes used in D059-D104:

CODED CAUSES -- READ ME FIRST (6 pages)
ICD - 09 CODES (153 pages)
ICD-09 RECODES (14 pages)
ICD - 10 CODES (197 pages)
ICD-10 RECODES (20 pages)

Section III: Linking Records

Using the 1968-2021 Mortality File in Conjunction with the Cross-Year Individual File

The 1968-2021 Mortality File is designed for linkage with the 1968-2021 cross-year Individual File. To link the two files, the analyst will need to match on the unique individual-specific identifiers. This unique identifier is a combination of two variables: "1968 Interview Number" (D002) and "Person Number" (D003). The corresponding variables for these unique identifiers on the cross-year Individual File are ER30001 and ER30002.

The 1968-2021 Mortality File is fully compatible only with the 1968-2021 Individual File. Nonmatches may occur if attempts are made to link it to prior cross-year Individual Files. Nonmatches are caused by changes in individual identifiers that may occur when corrections are made to cross-year files and when individuals move into a PSID household.

Individuals vary substantially in terms of which years they have been present in PSID family units over the course of the study. Information is available only for the years that the individual is present in a PSID family unit ("present" means living in the family unit or having left it to enter an institution). For more details about PSID tracking procedures and classification of people into family units, see the PSID User's Guide.

Using the 1968-2021 Mortality File with the Family Files

In order to access information from the Family Files, the Mortality File records must first be matched with the Individual File to obtain yearly identifiers (the Interview Number variables) for 1968-2021 Family data. This procedure is described in detail on our website. See the PSID website at https://psidonline.isr.umich.edu/ and choose 'Documentation' then 'File Structure.'