American Journal of Epidemiology

Vol. 170, No. 4
DOI: 10.1093/aje/kwp155
Advance Access publication June 30, 2009

Practice of Epidemiology

Identifying a National Death Index Match

Gerda G. Fillenbaum, Bruce M. Burchett, and Dan G. Blazer

Initially submitted December 17, 2008; accepted for publication May 11, 2009.

Data from the National Death Index (NDI) are frequently used to determine survival status in epidemiologic or clinical studies. On the basis of selected information submitted by the investigator, NDI returns a file containing a set of candidate matches. Although NDI deems some matches as perfect, multiple candidate matches may be available for other cases. Working across data from the Duke University site of the Established Populations for Epidemiologic Studies of the Elderly (EPESE), NDI, and the Social Security Death Index (SSDI), the authors found that, for this Established Populations for Epidemiologic Studies of the Elderly cohort of 1,896 cases born before 1922 and alive as of January 1, 1999, a match on Social Security number plus additional personal information (specific combinations of last name, first name, month of birth, day of birth) resulted in agreement between NDI and Social Security Death Index dates of death 94.7% of the time, while comparable agreement was found for only 12.3% of candidate decedents who did not have the required combination of information. Thus, an easy to apply algorithm facilitates accurate identification of NDI matches.

death certificates; epidemiologic methods; matching; mortality

Abbreviations: EPESE, Established Populations for Epidemiologic Studies of the Elderly; NDI, National Death Index; SSDI, Social Security Death Index.

For many outcome studies, it is important to know survival status and cause of death. The most valid information, available to investigators but not the general public, comes from the National Death Index (NDI) (1).

Having met NDI's conditions for access, investigators send to NDI as much information as possible on each case that matches NDI requirements. In return, NDI sends a file of candidate matches, which may include several potential matches for a given case. It is the applicant's responsibility to identify which of the potential matches are true matches.

Given the variable availability of usable data for the NDI match, there are many occasions when the information available for matching is limited. We describe an approach, based on our experience, to determining which limited set of information yields an acceptable match.

MATERIALS AND METHODS

Sample

The goal of our study was to determine survival status in the sample members of the Duke site of the Established Populations for Epidemiologic Studies of the Elderly (EPESE) (2). This is a 10-year study of a representative sample of community residents ($n = 4{,}162$) of 5 adjacent counties in North Carolina. All were born before 1922. They were 65–105 years of age when initially seen in 1986, were followed annually through 1992, and last interviewed in 1996. Survival status through 1998 had been established. This study was approved by the Duke University Medical Center Institutional Review Board.

National Death Index

The NDI is a compilation of regularly updated information from the death certificates held by the states. ''Available to investigators solely for statistical purposes in medical and health research'' and ''[n]ot accessible to organizations or the general public for legal, administrative, or genealogy purposes'' (1), the NDI is considered the ''gold standard'' for identification of death because of its source of data, and it has been reported to provide coverage of deaths superior to that of other sources (3). For matching purposes, NDI

Correspondence to Dr. Gerda G. Fillenbaum, Center for the Study of Aging and Human Development, Box 3003, Duke University Medical Center, Durham, NC 27710 (e-mail: ggf@geri.duke.edu).

requests a file containing as many as possible of the following elements: last name, first name, middle initial; father's surname; Social Security number; month, day, and year of birth; sex, race, marital status, state of residence, and state of birth.

Social Security Death Index

The Social Security Administration Death Master File, known as the Social Security Death Index (SSDI), is a publicly available database provided free online by some sites that, as of December 17, 2008, contained decedent information on 82,926,475 persons. Data come from state death certificates, reports of death from friends and relatives, funeral directors, financial institutions, and postal authorities (3–5). Because disbursement of Social Security funds is supposed to cease at death, there is concern that this information be accurate, although inaccuracies (over- and underinclusion of deaths) have been noted. The information available on the Death Master File includes first name, middle initial, and last name; day, month, and year of birth; day, month, and year of death; city, county, and state where income from Social Security was received before death (usually, but not necessarily, the place of death); Social Security number; and where the number was issued. SSDI provides no information on sex, race, or marital status. On the free-access SSDI site that we used (www.rootsweb.ancestry.com), any combination of Social Security number, first name, middle initial, and last name permits a search that, if successful, yields all the information listed above.

Data to be matched

In order to determine the survival status through 2006 of Duke EPESE sample members (cases) alive as of January 1, 1999, we sent to NDI a file of 1,896 cases that included information on last name and first name; Social Security number; month, day, and year of birth; sex; and race. Other information recommended by NDI was not available. Of the 1,896 cases, NDI rejected 22 and deemed 466 to be nonmatches, leaving 1,408 potentially matchable cases, not all of whom were necessarily dead; 3,505 potential candidate matches were provided. The data provided by NDI were state of death; month, day, and year of death; first and last names; which of the 9 digits of the Social Security number were matched; month and day of birth; age at death; sex, race, and marital status; state of residence; state of birth; whether a match was exact; and the number of possible matches for a given case. Some of the data provided could not be checked because Duke EPESE had not gathered the information (e.g., state of birth).

Procedures to identify a match

To determine which of the candidate records could be considered matches, we 1) ranked the information available according to our estimate of importance as a personal identifier, assigning weights in a manner that permitted rapid disaggregation of the summed score into its component elements, and 2) checked the resulting information against SSDI.

Ranking and weighting. We assigned the lowest ranked item a weight of 1 and doubled the weight with each succeeding item, as indicated: state of residence (weight = 1), marital status (weight = 2), race (weight = 4), day of birth (weight = 8), month of birth (weight = 16), first name (weight = 32), last name (weight = 64), and Social Security number (minimum of 7 digits in correct order; weight = 128). Sex was not included in our weighting system because we required sex to be an exact match. The weights were summed, yielding a unique score that can be readily disaggregated to indicate how it is constituted. For example, a score of 255 indicates that all the weighted characteristics match; a score of 244 indicates that day of birth, month of birth, first name, last name, and Social Security number match (sum of respective weights of 8, 16, 32, 64, and 128), while a score of 120 indicates that only day of birth, month of birth, first name, and last name match (sum of respective weights of 8, 16, 32, and 64). We sorted the sample so that entry into a group required personal identifiers of progressively less importance, while the upper end of the range *was 1 point less than* the *floor* of the immediately preceding group. For example, a match on first name, last name, and Social Security number was needed to enter the first group (lower bound = 224, indicated by "++" in Table 1). To form the second group (range, 198–223), we dropped the first name as a requirement, and additional characteristics marked by "+" in Table 1 were accepted. To determine their importance, we dropped day of birth and month of birth from the next group (range, 192–197). As a further test of day of birth and month of birth, the fourth group maintained these but dropped the last name, while the fifth group retained month of birth but dropped day of birth and excluded names. The sixth group is a test of Social Security number, the seventh group is a test of first and last names and date of birth, and the eighth and ninth groups are tests of progressively less information. We chose all cases in a group if the group was small or a random sample if the group was large (Table 1).

Checking against SSDI. Using the free, publicly accessible SSDI, we checked each person within each group. We first entered the Social Security number. If the Social Security number was not recognized by SSDI or seemed to apply to a different person, we entered last name and first name. Because many names are fairly common, entering last name and first name sometimes produced a large number of potential matches. Within each scoring range, we noted the accuracy with which SSDI identified each of the EPESE-based personal identifiers and NDI date of death, focusing in particular on year of death. On the basis of the percent correct in each of the selected samples, we extrapolated the number expected to be correctly identified in the original group. Separately for groups that were good matches (arbitrarily set at >80% match) and those that were poorer matches, the total number and, hence, the percent estimated to be deceased were determined.

RESULTS

Table 1 indicates the characteristics assessed in each NDI-based scoring range, the number of cases reviewed, and match agreement between EPESE and SSDI on day of birth and between NDI and SSDI on day of death. Good

Table 1. Weights, Ranging From 1 to 128, for National Death Index Information and Extent of Identification in the Social Security Death Index, Based on Survival Status Through 2006 of Duke EPESE Sample Members Alive in 1999[a,b]

State (Weight = 1)	Marital Status (Weight = 2)	Race (Weight = 4)	Day of Birth (Weight = 8)	Month of Birth (Weight = 16)	First Name (Weight = 32)	Last Name (Weight = 64)	Social Security Number (Weight = 128)	Score	Group Total, no.	No. Used	Exact Agreement, no.						Exact Deaths	
											Date of Birth			Date of Death				
											Day	Month	Year	Day	Month	Year	No.	%
+		+	+	+	++	++	++	224–255	872	26	22	25	25	24	25	25	25	96.2
+	++	++	+	+		++	++	198–223	132	20	14	15	12	18	19	17	17	85.0
+		+				++	++	192–197	5	5	2	2	2	1	3	2	2	40.0
+	+	+	++	++	+		++	152–191	28	15	13	14	11	13	14	14	14	93.3
++	+	+		++			++	129–151	9	9	3	3	3	0	1	0	0	00.0
							++	128	5	5	2	2	2	1	1	1	1	20.0
+	+	+	++	++	++	++		120–127	461	24	17	22	22	4	3	4	4	16.7
+	+	+	++ (–)	++	+	++		88–119	1,100	29	17	18	18	2	2	3	3	10.3
				+		+		<88	893	24	14	16	15	0	1	3	3	12.5

Abbreviation: EPESE, Established Populations for Epidemiologic Studies of the Elderly.

[a] "++" indicates characteristics that are necessary to reach the minimum score; "+" indicates additional characteristics needed to reach the top of the range, except for 88–119, where 1 characteristic (day of birth, indicated by "++ (–)"), necessary for the minimum score, is dropped and not used to reach the top of the range.

[b] Categories 120–127 and 88–95 each had 2 duplicated sample members.

agreement was found on values of 152 and over, *with the exception of values 192–197*. Of the 61 cases examined with scores of 152–191 or 198–255, 56 (92%) appear to be good matches. "Matches" not within these ranges are suspect and would require additional assessment. For our analyses, it is a given that there is a match on sex. In the group with scores of 198–255, only one case had a score of 198, and this person was not a match. The lowest score with a match was 215, suggesting that, in this category, an acceptable match requires Social Security number (weight = 128), last name (weight = 64), and month of birth (weight = 16) (i.e., the minimum score, 208). To attain a score in the 152–191 group range, cases required the minimum consisting of Social Security number (weight = 128), month of birth (weight = 16), and day of birth (weight = 8). The Social Security number alone (weight = 128) or with the last name (total score, 192) was not found to be a reliable match. Indeed, of the 152 cases individually checked, at least 9 Social Security numbers appeared to be inaccurate, with numbers transposed, or apparently those of a spouse, a parent, or an unrelated individual.

By using the best matching categories (score ranges, 152–191 and 198–255) and assuming that the randomly selected cases are representative, 1,032 cases would be identified as deceased without further checking. Of these, 977 or 94.7% appear to be recognized accurately. The remaining categories (n = 2,473 candidate matches) include 305 potentially deceased cases (12.3%).

DISCUSSION

Although the SSDI is not the "gold standard," it nevertheless offers certain information not provided by NDI (such as Social Security numbers for name matches), permitting a check on assumptions as to which combination of identifying characteristics results in an acceptable match.

The Social Security number was found to be a necessary, but not sufficient, requirement for a match. The addition of last name with minimal additional identifiers (as seen in group 192–197) was found to be inadequate. This group, however, was rare (5 instances out of 3,505 candidate matches). Possibly when the last name is present, other identifiers are present also.

The addition of last name, together with some other unique information (first name, day of birth, and month of birth), or Social Security number with day of birth and month of birth was, however, likely to provide a good match.

We found errors on 6% of the Social Security numbers evaluated. They included a transposition error and identification of a spouse, a parent, or somebody quite different. Our cohort was born before Social Security was instituted, some (in particular, women) may not have worked outside the home and so did not obtain personal Social Security numbers, and others may have worked in occupations that were not included in Social Security at that time. Some may be entitled to Medicare through a spouse and so present the spouse's number as their own. Such conditions create problems in identifying people by Social Security number alone. With certain exceptions (state employees in certain states,

railroad retirees), Social Security numbers are now required for all and are issued to newborns. Accurate matching for those born more recently should therefore be easier.

Additional issues may create problems for matching. We found that NDI, SSDI, and EPESE dates of birth did not always agree. It is unclear whether this reflects inaccuracy by EPESE interviewers, by persons completing the death certificate, or informant error. In none of these cases were source documents (birth certificates) or derived sources (driver's license, passport) consistently checked. If the SSDI day of birth derives from Social Security information, it may be the more accurate source, but it is unclear whether SSDI uses this source. So, interviewer error, subject error, and entry error may all be implicated, not just for dates of birth, but for the Social Security number also.

Nevertheless, parsimonious information (Social Security number plus additional personal information) could adequately distinguish survivors from decedents. In epidemiologic studies, the NDI is commonly used to identify survival status, date of death, and death certificate-determined cause of death, all matters of importance when planning health services for an aging society. Although matching is often mentioned, the criteria and their level of accuracy are rarely indicated. It is important that such matching be accurate, and it is helpful if the criteria used are readily available and easy to apply. We offer an approach that meets these requirements.

ACKNOWLEDGMENTS

Author affiliations: Center for the Study of Aging and Human Development, Duke University Medical Center, Durham, North Carolina (Gerda G. Fillenbaum, Dan G. Blazer, Bruce M. Burchett); Geriatric, Research, and Clinical Center, Veterans Administration Medical Center, Durham, North Carolina (Gerda G. Fillenbaum); and Department of Psychiatry and Behavioral Science, Duke University Medical Center, Durham, North Carolina (Dan G. Blazer).

The Duke site of the Established Populations for Epidemiologic Studies of the Elderly (EPESE) was funded by the National Institute on Aging (N01-AG12102, R01 AG12765, and R01 AG17559). Additional support was provided by grant 5P60 AG11268 (Claude D. Pepper Older Americans Independence Center, Duke University).

The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services.

Conflict of interest: none declared.

REFERENCES

1. National Center for Health Statistics. What is the NDI? Hyattsville, MD: Division of Vital Statistics, National Center for Health Statistics; 1999. (http://www.cdc.gov/nchs/r&d/ndi/what_is_ndi.htm). (Accessed June 18, 2009).
2. Cornoni-Huntley J, Blazer D, Lafferty M, et al. *Established Populations for Epidemiologic Studies of the Elderly: Resource Data Book.* Vol II. Washington, DC: Public Health Service, National Institutes of Health; 1990. (NIH publication no. 90-495).
3. Hill ME, Rosenwaike I. The Social Security Administration's Death Master File: the completeness of death reporting at older ages. *Soc Secur Bull.* 2001-2002;64(1):45–49.
4. Social Security Administration. Improving the usefulness of Social Security Administration's Death Master File. Baltimore, MD: Office of the Inspector General, Social Security Administration; 2000. (http://ssaonline.us/oig/ADOBEPDF/A-09-98-61011.pdf). (Accessed June 18, 2009).
5. Social Security Administration. Personally identifiable information made available to the general public via the Death Master File. Baltimore, MD: Office of the Inspector General, Social Security Administration; 2008. (http://www.ssa.gov/oig/ADOBEPDF/A-06-08-18042.pdf). (Accessed June 18, 2009).