All Records (Not) Being Equal: A Comparison of Coverage, Bias, and Utility for Five Mortality Record Sources

Celeste Stone, Sandy Eyster, Chris Plotts, Chris Bradley, Susan Lapham, and Leslie Scott
American Institutes for Research

Introduction

Determination of mortality status is an integral part of epidemiological and demographic research. Researchers currently have several alternatives for ascertaining mortality status, which include the Social Security Administration's (SSA) Death Master File (DMF), the National Center for Health Statistics' National Death Index (NDI), as well as consumer and commercial credit bureau databases. Internet sites, such as the Social Security Death Index (SSDI) and Ancestry.com are also available to researchers for this purpose. However, while the Internet's breadth and capacity has grown exponentially over the past 10 years, there has been little new research comparing the old standards (DMF; NDI) to new alternatives for ascertaining mortality status (e.g., information available on Internet websites). Additionally, major changes to the SSA's DMF scheduled to take place later this year are expected to reduce the number of death records available on the DMF by 36 percent *annually* (Social Security Administration, 2011). This paper provides the demographic and epidemiological research community with timely information about viable alternatives for ascertaining mortality status by examining the coverage rates and coverage biases of five mortality record sources.

Previous research has established that Internet searches provide a quick, inexpensive, and relatively accurate alternative to "gold standards" like the NDI (Schisterman & Whitcomb, 2004; Sesso, Paffenbarger, & Lee, 2000). Schisterman & Whitcomb's (2004) study of 374 decedents from the Patient Outcomes Study at Cedars Sinai Medical Center found "significant decreases in accuracy among foreign-born decedents, especially women, as well as among African-Americans." However, while research has established relationships between cognitive ability, personality, and survival (Friedman, 2000; Kern & Friedman, 2008; Roberts et al, 2007), no research to date has been able to go past assessing mortality record coverage bias in terms of basic demographics by using early life factors such as personality and intelligence. Therefore, research establishing relationships between cognitive ability, personality, and survival may be at risk of including uncorrected undercoverage biases.

Now more than ever, demographic and epidemiological researchers need up-to-date comparisons of the available methods for ascertaining morality status. The SSA has announced that it will cease disclosing state death records beginning on November 1, 2011. SSA anticipates that this change will result in a 36 percent reduction in the number of new death records available on the DMF annually and a 5 percent reduction in the number of historical death records available on the DMF (Social Security Administration, 2011). This change means that the DMF, which has to date been a reasonable alternative to the NDI and perhaps the best source of pre-1979 mortality, may have decreased utility in the very near future and researchers may need to turn to alternative sources in lieu of the DMF. But nobody has systematically evaluated the costs, availability, accuracy and bias of the alternatives?

We assess these factors using a 1960 cohort of 9th through 12th grade students for whom early life measures of knowledge, abilities, personality, and demographics are available. Using this nationally-representative cohort of persons currently between 62 and 76 years of age, this study systematically evaluates the coverage error and biases of five different mortality record sources on a number of key demographic factors, as well as early life cognitive ability and personality which have an established relationship with survival. In addition, this study evaluates the costs (e.g., dollars; person-hours) and flexibility (e.g., search identifiers available; batch vs. individual processing; information returned) to assess the relative utility of each method.

Methods

This study focuses on a subset of 613 individuals identified as deceased through any source during the tracking phase of a larger follow-up study of original Project Talent participants. Project Talent is a large, nationally representative longitudinal study of men and women who were in high school in 1960 and who, according to baseline records, were born between approximately 1942 and 1945. Conducted by the American Institutes for Research (AIR), Project Talent collected extensive information on characteristics and cognitive abilities of approximately 440,000 high school students in 1960. Subsequent follow-up studies collected data on the students' educational, career, and personal experiences through approximately age 30. In 2011, AIR partnered with the University of Michigan to begin the first in a series of follow-up studies. Tracking activities for the 2011 follow-up of a representative one percent subsample of the original Project Talent participants began in May 2011 and concluded in September 2011.

By the end of the tracking phase, 613 of the 4159 sampled participants were identified as deceased through one or more sources.[1] This paper focuses on the systematic review of the following five sources to determine the coverage rates (in terms of mortality status) using the 613 Project Talent 2011 follow-up decedents:

- Source 1: Social Security Death Administration's Death Master File (DMF)
- Source 2: National Death Index (NDI)
- Source 3: Consumer and commercial credit bureau database searches provided by Lexis Nexis (LN)
- Source 4: Social Security Death Index (available online) (SSDI)
- Source 5: Other Internet sites determined to be useful during participant tracking (ancestry.com; findagrave.com; thewall-usa.com; sortedbyname.com; obitiuaries.com; legacy.com; Google searches; local newspaper obituaries) (WEB)

Each of these sources offer slightly different information fields that can be used for locating decedents, including social security number, surname, given name, maiden name, date of death, date of birth, last know residence, and place of issuance. All available information will be used to assess the coverage of each source. The DMF and LN will first be searched using batch data matching, followed by intensive individual record searches for records not found during the batch searches. All records will be submitted to the NDI as decedents. SSDI's online database is not downloadable and while there are

[1] Sources included DMF and NDI matches; household and classmate reports; and Accurint and SSDI searches.

software packages available for searching SSDI's online database, systematic searches of the Social Security Death Index will be carried out individually. Finally, we will conduct systematic individual searches of seven specific websites and one search engine (Google); local newspaper obituaries that are available online will be searched as a last resort. For each source, record will be classified as definitely deceased, possibly deceased, or presumed alive. Time data will be collected and recorded for all individual searches.

Because the NDI database is limited to deaths occurring since 1979 and through 2008, two coverage rates will be assessed for each source: deaths occurring from 1979 to 2008 (conditional coverage) and deaths occurring from 1960 to August 2011 (overall coverage). To assess the amount of coverage bias for each of the five mortality record sources, we compare the distributions for the group of people identified as "definitely deceased" across each of the five sources for a number of different demographic characteristics including, race, sex, year of birth, age at death, and year of death. We also compare the distributions on a number of different personality and cognitive measures shown to be associated with mortality, as well as academic performance, self-reported family socioeconomic status, and several measures of self-reported health. Continuous variables are expressed as a mean value and categorical values expressed as a percentage, with differences compared using two-tailed t-tests and chi-square statistics, respectively. Multivariate logistic regression models will be used to produce adjusted odds ratios that person will be missed if a given mortality record source is used.

The measures evaluated in this analysis come primarily from the 1960 base-year collection. Gender was self-reported by students. Cognitive ability is assessed through multiple scales, including: Vocabulary I (crystallized intelligence); Spatial visualization in 3 dimensions (visualization); Mathematics II (Quantitative ability); Clerical checking (speed); Memory for sentences (Long-term retrieval); Memory for words (Short-term retrieval), Abstract reasoning; Arithmetic reasoning (Mathematical reasoning); and Creativity. Personality is assessed through 10 scales: Sociability; Social Sensitivity; Impulsiveness; Vigor; Calmness; Tidiness; Culture; Leadership; Self-Confidence; and Mature Personality. Race was not directly measured in the base year collection, but was collected in the 5- and 11-year follow-ups; therefore we have self-reported race of respondent for approximately 50% of respondents. For those cases without a direct measure of race, we will use information about the distribution of race/ethnicity within the school as a proxy for race.[2] Academic performance is measured by self-reported grades across specific courses (mathematics, science, foreign languages, history and social studies, English, vocational, and business or commercial) and in all courses. Respondent health background measures from the base year include self reported health prior to age 10 and in the past three years; number of days sick in bed in the past year; number of treatments by a doctor in the past 6 months; several measures of self-reported health issues/disabilities (e.g., need for corrective eyeglasses, use of hearing aids, speech impediment, normal use of legs, arms, hands, ever having been knocked unconscious, etc.). Family socioeconomic status in 1960 will be measured using a composite of four variables reported by

[2] Given the highly segregated nature of schools in 1960, information about the distribution of race within schools is a reasonable proxy for race of respondents.

the students in 1960: number of books in home, number of rooms in home, student-reported financial well-being.

Anticipated Results

These analyses will contribute to mortality research by estimating the coverage and coverage bias of five different mortality record sources. Consistent with previous research, we expect that our findings will show Internet sources to be viable alternatives to the DMF and NDI, particularly where key identifiers (e.g., SSN) are not available, but at a sacrifice of time. In terms of coverage, the SSDI and DMF are likely to produce very similar results. LN will likely have the least amount of coverage for decedents for whom key identifiers are not available. Compared to the DMF and SSDI, we expect NDI and WEB sources to be better for determining the morality status of women. While we expect the NDI to have the best coverage and least bias for deaths occurring between 1979 and through 2008, we expect the NDI to have greater differences on personality measures like impulsive behavior, which might be associated with deaths occurring before 1979. Finally, while we expect WEB alone to have lower coverage overall, we also expect it to be the best for finding particular types of mortality records (e.g., war-related, certain accidental deaths, deaths in certain time periods).

References

Friedman, H.S. (2000). Long-term relations of personality and health: Dynamisms, mechanisms, tropisms. *Journal of Personality, 68*, 1089-1108.

Kern, M.L., & Friedman, H. S. (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology, 27*(5), 505-512.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A.,, and Goldberg, L.R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science.* 2(4):313-345.

Schisterman, E. & Whitcomb, B. (2004). Use of the Social Security Administration Death Master File for ascertainment of mortality status. *Population Health Metrics,* 2(2).

Sesso, H.D., Paffenbarger, R.S., & Lee, I.M. (2000). Comparison of National Death Index and World Wide Web death searches. *American Journal of Epidemiology 152*(2), 107-11.

Social Security Administration. (2011, August 24). *Fact sheet-change to the public Death Master File* [Online availability of SSA's Death Master File]. Retrieved from http://ssa-custhelp.ssa.gov/app/answers/detail/a_id/149/~/online-availability-of-ssa%27s-death-master-file