Annals of Epidemiology xxx (2016) 1–7

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Do inferences about mortality rates and disparities vary by source of mortality information?

John Robert Warren PhD [a,*], Carolina Milesi PhD [b], Karen Grigorian MPM [b], Melissa Humphries PhD [c], Chandra Muller PhD [d], Eric Grodsky PhD [e]

[a] Department of Sociology, University of Minnesota, Minneapolis
[b] NORC at the University of Chicago, Chicago, IL
[c] Texas Higher Education Coordinating Board, Austin
[d] Department of Sociology, University of Texas-Austin, Austin
[e] Department of Sociology, University of Wisconsin-Madison, Madison

ARTICLE INFO

Article history:
Received 13 June 2016
Accepted 3 November 2016
Available online xxx

Keywords:
Survey methods
Death records
Differential mortality
Mortality rate

ABSTRACT

Purpose: Researchers who study mortality among survey participants have multiple options for obtaining information about which participants died (and when and how they died). Some use public record and commercial databases; others use the National Death Index; some use the Social Security Death Master File; and still others triangulate sources and use Internet searches and genealogic methods. We ask how inferences about mortality rates and disparities depend on the choice of source of mortality information.
Methods: Using data on a large, nationally representative cohort of people who were first interviewed as high school sophomores in 1980 and for whom we have extensive identifying information, we describe mortality rates and disparities through about age 50 using four separate sources of mortality data. We rely on cross-tabular and multivariate logistic regression models.
Results: These sources of mortality information often disagree about which of our panelists died by about age 50 and also about overall mortality rates. However, differences in mortality rates (i.e., by sex, race/ethnicity, education) are similar across of sources of mortality data.
Conclusion: Researchers' source of mortality information affects estimates of overall mortality rates but not estimates of differential mortality by sex, race and/or ethnicity, or education.

Introduction

Research on mortality is often based on survey data that have been linked to administrative or public records on sample members' deaths. The National Death Index (NDI), the Social Security Death Master File (SSDMF), commercial databases (e.g., LexisNexis' Accurint database), and Web-based genealogical resources all provide such information. These resources have different strengths and weaknesses regarding population coverage, timeliness of updates, and accuracy. Most researchers seem to assume that the results of their analyses are unaffected by their source of mortality information.

How do inferences about mortality rates and disparities—at least through midlife—depend on the source of mortality information? Using data on a nationally representative sample of Americans first interviewed in high school in 1980 and for whom we have extensive identifying information, we describe mortality rates and disparities through age 50 using four sources of mortality data. Because we study the same subjects using the same measures of those subjects' characteristics, differences across analyses in inferences about mortality rates and disparities are entirely due to differences across sources of mortality data.

There are two major reasons to measure survey respondents' subsequent mortality. First, there are applied and academic research reasons for better understanding mortality by studying postsurvey death for example [1,2]. Second, researchers conducting longitudinal surveys need to know which sample members are deceased so they accurately report response rates and do not invest scarce resources attempting to locate them or convincing them to participate in future surveys. Most researchers make use of the SSDMF or NDI for these purposes.

The SSDMF contains information about people who had Social Security numbers and whose deaths were reported to the Social

* Corresponding author. Department of Sociology, University of Minnesota, 909 Social Sciences, 267 19th Ave S., Minneapolis, MN 55455. Tel.: 612 624 2310; fax: 612 624 7020.
E-mail address: warre046@umn.edu (J.R. Warren).

http://dx.doi.org/10.1016/j.annepidem.2016.11.003
1047-2797/

Security Administration in 1962 or later. Mortality information is based on reports from funeral directors, family members, financial institutions, and government agencies. Coverage of U.S. deaths in the SSDMF is generally between 85% and 90% but varies by age, sex, race and/or ethnicity, and other personal attributes [3–7]. Conversely, because of sharing social security numbers and data entry errors, some individuals appear in the SSDMF who are not deceased.

The NDI is a centralized database of information from states' vital statistics offices [8]. States gather death certificate information from physicians, funeral directors, and local vital statistics and health departments. Although coverage rates in the NDI are typically higher than in the SSDMF [9–11], deaths before 1979 are not recorded. Also, recording errors by proxy informants (e.g., physicians, funeral directors) can harm the quality of the identifying information in the NDI. Finally, whereas access to the SSDMF is free and immediately available, the NDI can be a major expense for large surveys and the submission process can be time consuming. Currently, NDI charges $0.21 for each person-year of records search. A survey of 10,000 people observed over 10 years would thus need to pay $10{,}000 \times 10 \times 0.21 = \$21{,}000$ to determine panelists' mortality status.

Previous efforts to assess bias and error in the NDI and SSDMF have followed two approaches. First, lacking a "gold standard" against which to validate NDI and SSDMF information, some researchers compare the rate at which they agree regarding people's mortality status, for example [9,10,12]; we follow this approach. Second, other researchers compare NDI and SSDMF mortality information using samples of people known to be dead, for example [3–6,11,13,14]. Such efforts show that the accuracy of these resources varies by age, race and/or ethnicity, gender, and other attributes, for example [3,5,6,9,13,14]. Accuracy also depends on the quality of identifying information about sample members, for example [14–16]. No prior research has asked how inferences about socioeconomic and demographic disparities in mortality depend on the source of mortality data.

We compare mortality rates and disparities across the NDI, the SSDMF, Accurint, and searches that seek to triangulate information across multiple sources. One goal is to draw conclusions about how researchers ought to ascertain sample members' mortality. Another is to understand whether conclusions about mortality disparities depend on the source of mortality data.

Material and methods

High School & Beyond (HSB) began in 1980 with a nationally representative random sample of 30,030 sophomores and 28,240 seniors in 1020 high schools; all of our sample sizes are rounded to the nearest 10 per the requirements of our restricted data use agreement. In 1980, sampled students completed an in-school survey, which included questions about social background and demographic characteristics; questions about schooling experiences, activities, and plans; and achievement tests in multiple subjects.

A random sample of 14,830 HSB sophomores and 12,000 HSB seniors were included in longitudinal follow-up surveys. Sophomores were reinterviewed in 1982, 1984, 1986, and 1992; seniors were reinterviewed in 1982, 1984, and 1986. Survey data were supplemented with secondary transcripts for sophomores and post-secondary transcripts for both sophomores and seniors. In 2014, we completed a new round of data collection with HSB sophomores, who were then about 50 years old. We attempted to locate and interview all of the 14,830 sophomores who were included in the longitudinal component of HSB after 1980, regardless of whether they participated in any previous wave of the survey or were previously recorded as deceased. The 2014 surveys gathered basic information about post-1992 educational attainments; family circumstances; health and functional limitations; and labor force activities. We obtained completed surveys from 8790 panel members.

There are three reasons we were unable to complete surveys with the other 6040 (14,830−8790) HSB sophomores. First, despite having access to sample members' identifying information, we were unable to locate some living sample members. Second, some living sample members were located but were unwilling or unable to complete surveys. Third, some sample members died by 2014.

We are confident that sample members were alive in 2014 if (a) we interviewed them in 2014 or (b) we located them but were unable to obtain completed interviews. The 2014 survey began with a set of questions designed to ensure that we interviewed only correct sample members. We only classified sample members as "located" when we communicated directly with them or a trustworthy proxy informant who indicated that we had located the correct (still living) person. However, we cannot distinguish between sample members who (a) died before 2014 or (b) were alive in 2014 but could not be located. Thus, we cannot accurately calculate statistics like survey response rates and we cannot know which sample members to recontact in future surveys. To do those things, we need to know which sample members were deceased in 2014.

We restrict our focus to the 5470 HSB sophomores who did not complete surveys in 2014 but who did respond to the 1980 HSB survey; the latter is necessary for obtaining variables predictive of subsequent mortality. To handle the small amounts of item-level missing data on 1980 measures (see Table 1), we use Stata's (StataCorp, College Station, TX) ICE routine for multiple imputation (and generate five multiply imputed data sets that are then combined using appropriate Stata routines). All of our analyses are weighted by the 1980 base year sampling weight.

Almost all of the social, demographic, and educational variables we use to predict mortality are derived from the 1980 survey; a few (such as completed years of schooling) were obtained from the 1982–1992 surveys or transcript data. Those measures are described in Table 1 and include measures of educational attainment, gender, race and/or ethnicity, nativity, socioeconomic background, reading and math achievement, high school grade point average, self-concept, locus of control, body mass index based on separate questions about height and weight, and whether sophomores had physical conditions that limited their activities. If our objective were to conduct rigorous analyses of the social, demographic, or other predictors of mortality, we would include more (and more theoretically motivated) measures. However, this set of measures will serve to describe the ways in which broad contours of conclusions about predictors of mortality depend on how mortality status is ascertained.

We used four methods to ascertain mortality status. All four methods enumerate deaths that occurred from 1980 onward, although the latest available information differs slightly across sources of information (mainly because of the 2-year lag in the available NDI data). Because the vast majority of our panelists were born in 1964 or 1965, we describe mortality through midlife (i.e., about age 50).

Accurint measure

Before the 2014 survey, we used Accurint—a fee-based subscription service that compiles public records and information from private companies and credit bureaus. As shown in Table 2, there were 470 sample members who were listed as deceased in batch and/or individual searches of the Accurint data.