Do Inferences about Mortality Rates and Disparities Vary by Source of Mortality Information?

October 2016

John Robert Warren (Contact Author) ^a

Carolina Milesi ^b

Karen Grigorian ^b

Melissa Humphries ^c

Chandra Muller ^d

Eric Grodsky ^e

^a University of Minnesota, Department of Sociology, 909 Social Sciences, 267 19th Ave S., Minneapolis, MN 55455, warre046@umn.edu

^b NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603

^c Texas Higher Education Coordinating Board, 1200 E. Anderson Lane, Austin, TX 78752

^d University of Texas-Austin, Department of Sociology, 116 Inner Campus Dr Stop G6000, Austin, TX 78712

^e University of Wisconsin-Madison, Department of Sociology, 8128 William H. Sewell Social Sciences Building, 1180 Observatory Drive, Madison, WI 53706-1393

Do Inferences about Mortality Rates and Disparities Vary by Source of Mortality Information?

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/ethnicity, or education.

KEYWORDS

Survey Methods; Death Records; Differential Mortality; Mortality Rate

LIST OF ABBREVIATIONS

HSB = High School & Beyond

IES = Institute on Education Sciences

NCHS = National Center for Health Statistics

NDI = National Death Index

SSDMF = Social Security Death Master File

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 post-survey death (e.g., 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 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/ethnicity, and other personal attributes (3-7). Conversely, because of such as 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 prior to 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 (e.g., 9, 10, 12); we follow this approach. Second, other researchers compare NDI and SSDMF mortality information using samples of people known to be dead (e.g., 3, 4-6,

11, 13, 14). Such efforts show that the accuracy of these resources varies by age, race/ethnicity, gender, and other attributes (e.g., 3, 5, 6, 9, 13, 14). Accuracy also depends on the quality of identifying information about sample members (e.g., 14, 15, 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 1,020 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 re-interviewed in 1982, 1984, 1986, and 1992; seniors were re-interviewed 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 8,790 panel members.

There are three reasons we were unable to complete surveys with the other 6,040 (14,830 – 8,790) 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 prior to 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 re-contact

in future surveys. To do those things, we need to know which sample members were deceased in 2014.

We restrict our focus to the 5,470 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 employ Stata's 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; 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/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 utilized 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 two-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. Prior to 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.

SSDMF Measure. After the 2014 survey, we used the linking algorithm RECLINK in Stata to perform probabilistic linkages to the SSDMF. We then visually inspected records and considered the plausibility of the matches suggested by RECLINK. As shown in Table 2, these procedures suggested that 330 sample members were deceased.

NDI Measure. NDI uses a probabilistic linking algorithm (8) to identify a possible set of matches for each record submitted for each person. Then, possible matches that exceed a pre-determined threshold are declared by NDI to be a valid match. As shown in Table 2, NDI declared 600 valid matches to HSB sophomores.

NORC Measure. Prior to and during the 2014 survey, NORC conducted a manual mortality search that focused on investigating and verifying mortality status for sample members who could not be located. NORC began with Accurint® (but not SSDI or NDI) records, and then triangulated across a variety of online services, obituary and grave listings, genealogical websites, and phone calls to key informants. 600 sample members were inferred to be deceased via this method.

RESULTS

Table 2 presents the number and percentage dead among the 5,470 sophomore sample members who responded in 1980 but not in 2014, separately by source of mortality information. The SSDMF (6.0%) and Accurint® (8.6%) indicate relatively fewer deaths; NORC and NDI each indicate that about 11% of these 5,470 were deceased. Of the 5,470 panelists, 13.3% were deceased according to at least one source of information but only 5.1% were deceased according to all of them. How do these mortality rates at midlife compare to what we would expect for this cohort of Americans?

First, using life table estimates—based on the all-race, all-sex rate of survival from age 15-20 in 1980 to age 20-25 in 1985, from age 20-25 in 1985 to age 25-30 in 1990, etc. (e.g., 17, 18)—we calculate an expected mortality rate of 5.9%. Thus we should expect $14,830 \times 0.059 = 880$ deceased sample members based on life table estimates. Second, dividing (a) the weighted number of people in the American Community Survey in 2013 who were born between 1963 and 1965 and lived in the U.S. in 1980 by (b) the weighted number of people in the 1980 U.S. Census who were born between 1963 and 1965 we calculate an expected mortality rate of 4.8%. Thus we should expect $14,830 \times 0.048 = 710$ deceased sample members based on Census estimates. Across the four sources of mortality information, Table 2 shows that between 330 and 600 sample members are deceased; this implies that each of the four sources are missing a large number of deaths. Whether these undercounts result from under-coverage in the four mortality databases, coverage issues from the sample design which only includes those that who (a) attended high school, and (b) those who were U.S. residents in 1980, or from problems with record linkage is unclear.

Regardless of the overall rates of death implied by each source of mortality information, do the data sources agree about which sample members died? In Table 3, we describe consistency in mortality status across source of mortality information. Most (86%) of the 5,470 sample members who did not respond to the 2014 survey are consistently classified as living in 2014. Of those *ever* classified as dead by any source of information, only about a third are consistently classified as dead.

In Table 4 we cross-classify mortality status for each combination of sources of information about mortality. For each cross-tabulation, we report both overall rates of agreement in classification and Cohen's (19) kappa, a measure of inter-rater reliability that logically ranges from -1 (complete disagreement) to +1 (complete agreement). Rates of agreement are always at least 94%, but this is because most respondents are not dead according to any data source. Thus, we rely mainly on kappa, which calibrates rates of agreement by accounting for the level of agreement expected by chance (which is high in this case). Across the six cross-classifications, kappa ranges from 0.61 to 0.88, indicating only moderately high levels of agreement. Kappa is lowest for any of the cross-classifications involving SSDMF.

Even disagreement across data sources about which sample members are deceased does not necessarily mean that inferences about mortality disparities will be affected by choice of information source about mortality status. In Table 5, we report sample members' mortality status separately by several of the individual-level attributes described in Table 1. As expected, we observe that mortality declines as educational attainment increases; that mortality is higher for men than for women; and that mortality is a bit lower for non-Hispanic Whites. Generally

speaking, however, descriptions about mortality differentials do not depend on which source of information we use to classify mortality status. Overall mortality rates are higher when we rely on NORC's or NDI's mortality information, but disparities and differentials in mortality look quite similar.

Substantive analyses of mortality disparities would probably use multivariate regression techniques to understand the ways in which mortality status varies by socioeconomic, demographic, and other circumstances. In Table 6 we report results of logistic regression models in which mortality status (0 equals living, 1 equals deceased) is a function of all of the predictors described in Table 1. We estimate separate models using mortality information from Accurint®, SSDMF, NDI, and NORC.

In general, the results in Table 6 resemble results from more substantively-motivated analyses of the social and demographic predictors of early mortality. Women are always significantly less likely to have died by about age 50, although point estimates describing the magnitude of that conditional association vary a bit. All else constant, sample members' odds of early mortality are not significantly related to any of the other social, economic, or demographic predictors in the model. However, as shown in Table 6, coefficients are generally in expected direction. We see little or no evidence that conditional associations between any of the predictor variables and mortality differ substantially depending on the data we used to determine sample members' mortality status.

DISCUSSION

Whether someone is alive or dead would seem to be among the easiest things to measure about them. However, when information about mortality status gets reported by fallible human beings and then aggregated up to form complex administrative databases, there is room for error. Different processes may contribute to variance in mortality in different data sets. For example, the triggering event for NDI is a death reported to state vital statistics while sometimes the triggering event for SSMDf is a claim by a beneficiary. The NDI, SSMDf, and other mortality data resources do not capture all deaths in the United States, and their records contain inaccuracies that can reduce the number and quality of matches to survey data. At the same time, survey data themselves are flawed: The identifying information in them also originates from fallible human beings, and so errors are to be expected.

On top of all of this, not all Americans are well connected to administrative data and surveillance systems. Some Americans—perhaps those most likely to die before midlife—actively seek to avoid the attention of financial institutions, state agencies, and other bureaucratic institutions. For example, Brayne (20: 367) describes “‘system avoidance,’ whereby individuals who have had contact with the criminal justice system avoid surveilling institutions that keep formal records.” People’s ability to avoid systematic surveillance likely varies across demographic and socioeconomic groups.

Prior research has documented rates of coverage of the NDI, SSMDf, and other resources. We know that the NDI has more complete coverage than the SSMDf, although the rate of coverage varies across demographic groups. In this paper we ask two related questions: Do our conclusions about disparities in mortality through midlife depend on our source of mortality information? Also, how should researchers determine mortality status?

We found that our sources of mortality information produce different mortality rates, at least through about age 50. The mortality rate implied by the SSDMF was lowest; this is not surprising given previous evidence that the SSDMF reflects fewer deaths among young people and racial/ethnic minorities and the age and diversity of our sample.

However, our conclusions about disparities in mortality through midlife are quite similar across source of mortality information. In bivariate analyses (Table 5), we saw disparities in mortality by gender, education, and socioeconomic background even at this relatively young age; these disparities were similar across sources of mortality information. In multivariate analyses (Table 6) we found that only gender was a significant predictor of mortality; again, these disparities were similar across sources of mortality information. Whether we would come to similar conclusions about mortality at older ages—or about disparities in timing or cause of death—remains to be seen.

A more basic point should not be overlooked: All four sources of mortality information appear to be undercounting deaths—perhaps dramatically. Based on life tables and U.S. Census data, we would have expected between 710 and 880 deaths. We showed in Table 2 that none of the data sources identified more than 600 sample members as dead. This is not a trivial undercount. So while our inferences about inequalities in mortality may not be affected by our source of information about deaths, our inferences about mortality rates almost certainly are. This undercount may be a function of limitations in the sources of mortality information, of record linkage failures, and/or of limitations of the HSB survey design (e.g., selective panel attrition).

How should survey researchers ascertain the mortality status of sample members? No one source of information is sufficient. Each has systematic limitations, but each also may include idiosyncratic errors. Research projects should not rely exclusively on NDI because of the two-year delay in data availability and because of uncertainty inherent in the probabilistic linking algorithm (not to mention the expense). They also should not rely exclusively on the SSDMF or databases like Accurint® because of issues of under-coverage.

We advocate a creative integration of as many data sources as possible—something like the measure we label as NORC above. We began with relatively inexpensive Accurint® and SSDMF searches. We then did an independent search of NDI; had our NDI search been dependent on the results of the Accurint® or SSDMF searches we would likely have missed many deaths. Then, and most importantly, we treated the apparent matches in Accurint®, SSDMF, and NDI as *suggestions* of mortality, not as confirmation of it. For all sample members deceased according to any of these resources, we conducted follow-up investigation using Internet, genealogical resources, and outreach to informants.

Data quality issues are such that we will rarely be 100% certain that a non-responding sample member is deceased. However, by triangulating information across multiple sources we can rule out false positives, have greater confidence about which sample members are actually deceased, and better understand the strengths and limitations of survey data. At the same time, it seems to be the case that substantive conclusions about the correlates and predictors of mortality are fairly robust to these methodological concerns.

ACKNOWLEDGEMENTS

The 2014 High School & Beyond project was supported by the Alfred P. Sloan Foundation (Grant 2012-10-27), the Institute for Education Sciences of the U.S. Department of Education (Grant R305U140001), and the National Science Foundation (Grants HRD1348527 and HRD1348557). This project also benefited from support provided by the Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD) to the University of Texas at Austin (R24-HD042849), the University of Wisconsin-Madison (P2C-HD047873), and the University of Minnesota (P2C-HH041023), as well as direct funding from NORC at the University of Chicago. We are also very grateful for constructive suggestions by Jack DeWaard, Robert Hummer, and John Stevenson. However, errors or omissions are entirely the responsibility of the authors. Note that this manuscript has been subject to disclosure review and has been approved by the U.S. Department of Education's Institute for Education Sciences in line with the terms of the HS&B restricted use data agreement.

REFERENCES

1. Elo IT. Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective. *Annual Review of Sociology*. 2009;35:553-72. PubMed PMID: WOS:000270116800026. English.
2. Montez JK, Berkman LF. Trends in the Educational Gradient of Mortality Among US Adults Aged 45 to 84 Years: Bringing Regional Context Into the Explanation. *American Journal of Public Health*. 2014 Jan;104(1):E82-E90. PubMed PMID: WOS:000341701400016. English.

3. Ashley T, Cheung L, Wokanovicz R. Accuracy of vital status ascertainment using the Social Security Death Master File in a deceased population. *Journal of insurance medicine*. 2012;43(3):135-44. PubMed PMID: 23451614.
4. Hauser TH, Ho KK. Accuracy of on-line databases in determining vital status. *J Clin Epidemiol*. 2001 Dec;54(12):1267-70. PubMed PMID: 11750196.
5. Huntington JT, Butterfield M, Fisher J, Torrent D, Bloomston M. The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients. *Am J Cancer Res*. 2013;3(5):518-22. PubMed PMID: WOS:000331069500008. English.
6. Schisterman EF, Whitcomb BW. Use of the Social Security Administration Death Master File for ascertainment of mortality status. *Population Health Metrics*. 2004;2(1).
7. Hill ME, Rosenwaike I. The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages *Social Security Bulletin*. 2001/2002;64:45-51.
8. National Center for Health Statistics. *National Death Index User's Guide*. Hyattsville, MD: National Center for Health Statistics; 2013.
9. Hanna DB, Pfeiffer MR, Sackoff JE, Selik RM, Begier EM, Torian LV. Comparing the National Death Index and the Social Security Administration's Death Master File to Ascertain Death in HIV Surveillance. *Public Health Reports*. 2009 Nov-Dec;124(6):850-60. PubMed PMID: WOS:000270608900013. English.
10. Lawler TP, Lawler FH. Mortality status of professional basketball players: sensitivity and concordance of four web-based sources. *International Journal of Academic Research*. 2011;3(2):589-91.
11. Stampfer MJ, Willett WC, Speizer FE, Dysert DC, Lipnick R, Rosner B, et al. Test of the National Death Index. *American Journal of Epidemiology*. 1984;119(5):837-9. PubMed PMID: WOS:A1984SR57500018. English.

12. Sesso HD, Paffenbarger RS, Lee IM. Comparison of National Death Index and World Wide Web death searches. *American Journal of Epidemiology*. 2000 Jul 15;152(2):107-11. PubMed PMID: WOS:000088211500002. English.
13. Boyle CA, Decoufle P. National Sources of Vital Status Information - Extent of Coverage and Possible Selectivity in Reporting. *American Journal of Epidemiology*. 1990 Jan;131(1):160-8. PubMed PMID: WOS:A1990CE99100017. English.
14. Curb JD, Ford CE, Pressel S, Palmer M, Babcock C, Hawkins CM. Ascertainment of Vital Status through the National Death Index and the Social-Security-Administration. *American Journal of Epidemiology*. 1985;121(5):754-66. PubMed PMID: WOS:A1985AGA3900015. English.
15. Lash TL, Silliman RA. A comparison of the national death index and social security administration databases to ascertain vital status. *Epidemiology*. 2001 Mar;12(2):259-61. PubMed PMID: WOS:000167139400021. English.
16. Williams BC, Demitrack LB, Fries BE. The Accuracy of the National Death Index When Personal Identifiers Other Than Social-Security Number Are Used. *American Journal of Public Health*. 1992 Aug;82(8):1145-7. PubMed PMID: WOS:A1992JF31000017. English.
17. National Center for Health Statistics. Vital Statistics of the United States, 1980, Vol. 11, Sec. 6, Life Tables. DHHS Pub. No. (PHS) 84-1104. Washington, D.C.: U.S. Government Printing Office; 1984.
18. National Center for Health Statistics. Vital Statistics of the United States, 1985, Vol. 11, Sec. 6, Life Tables. DHHS Pub. No. (PHS) 88-1104. Washington, D.C.: U.S. Government Printing Office; 1988.
19. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20(1):37-46. PubMed PMID: WOS:A1960CCC3600004. English.
20. Brayne S. Surveillance and System Avoidance: Criminal Justice Contact and Institutional Attachment. *Am Sociol Rev*. 2014 Jun;79(3):367-91. PubMed PMID: WOS:000340424400001. English.

Table 1. Descriptive Statistics for Predictor Variables Before and After Imputation

	Prior to Imputation					After Imputation (n=5,470)			
	Mean / %	(sd)	n	Min.	Max.	Mean / %	SD	Min	Max
Educational Attainment									
Less than High School	7.6%		4,340	0.0	1.0	8.4%		0.0	1.0
High School Graduate	36.3%		4,340	0.0	1.0	36.5%		0.0	1.0
Some College, No BA	41.0%		4,340	0.0	1.0	40.4%		0.0	1.0
BA or Higher	15.1%		4,340	0.0	1.0	14.7%		0.0	1.0
Gender									
Male	54.9%		5,470	0.0	1.0	54.9%		0.0	1.0
Female	45.1%		5,470	0.0	1.0	45.1%		0.0	1.0
Race/Ethnicity									
White	64.8%		5,470	0.0	1.0	64.8%		0.0	1.0
Hispanic	16.2%		5,470	0.0	1.0	16.2%		0.0	1.0
Black	15.8%		5,470	0.0	1.0	15.8%		0.0	1.0
All Others	3.2%		5,470	0.0	1.0	3.2%		0.0	1.0
Nativity									
Born in United States	94.1%		5,330	0.0	1.0	94.1%		0.0	1.0
Born Abroad	5.9%		5,330	0.0	1.0	5.9%		0.0	1.0
Father's Educational Attainment									
Less than High School	28.4%		3,450	0.0	1.0	30.1%		0.0	1.0
High School Graduate	31.5%		3,450	0.0	1.0	32.0%		0.0	1.0
Some College or More	40.1%		3,450	0.0	1.0	37.9%		0.0	1.0
Mother's Educational Attainment									
Less than High School	23.2%		4,060	0.0	1.0	24.5%		0.0	1.0
High School Graduate	45.4%		4,060	0.0	1.0	44.9%		0.0	1.0
Some College or More	31.5%		4,060	0.0	1.0	30.6%		0.0	1.0
Academic Achievement in 1980									
Reading Test Score	5.7	(4.5)	4,860	-1.3	19.0	7.0	(5.8)	-1.3	19.0
Math Test Score	9.8	(9.3)	4,870	-4.5	38.0	10.9	(10.5)	-4.5	38.0
Self-reported GPA	2.5	(0.8)	5,390	0.0	4.0	2.5	(0.8)	0.0	4.0
Non-Cognitive Skills in 1980									
Self-concept	-0.1	(1.0)	5,140	-4.8	1.7	-0.1	(1.0)	-4.8	1.7
Locus of control	-0.2	(1.0)	5,120	-3.7	2.2	-0.2	(1.0)	-3.7	2.2
Work orientation	0.0	(1.1)	5,180	-6.3	1.7	0.0	(1.1)	-6.3	1.7
Health and Disability in 1980									
Body Mass Index	21.3	(3.7)	4,850	7.8	122.6	21.3	(3.7)	7.8	122.6
Limiting Physical Condition	10.2%		4,960	0.0	1.0	10.5%		0.0	1.0

Note : Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Analyses are weighted by the base year sampling weight BYWT. Imputations are performed using chained equations in Stata's ice command. Sample sizes are rounded as per IES Restricted Data security requirements.

Table 2. Descriptive Statistics for Mortality Outcomes

	Living	Deceased	% Deceased
Accurant®	5,000	470	8.6%
Social Security Death Master File	5,140	330	6.0%
National Death Index	4,870	600	11.0%
NORC's Final Disposition After 2014 Locating Effort	4,870	600	11.0%
Deceased According to Any of the Four Sources	4,740	730	13.3%
Deceased According to All Four Sources	5,190	280	5.1%

Note : Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Analyses are not weighted. Sample sizes are rounded as per IES Restricted Data security requirements.

Table 3. Consistency in Mortality Status across Sources of Information

	Accurint©	SSDMF	NDI	NORC	n	%
Deceased according to:						
FOUR Sources	Deceased	Deceased	Deceased	Deceased	280	5%
THREE Sources	Deceased	Living	Deceased	Deceased	100	2%
	Deceased	Deceased	Living	Deceased	20	0%
	Living	Deceased	Deceased	Deceased	10	0%
	Deceased	Deceased	Deceased	Living	*	0%
TWO Sources	Living	Living	Deceased	Deceased	140	3%
	Deceased	Living	Living	Deceased	10	0%
	Living	Deceased	Living	Deceased	*	0%
	Deceased	Living	Deceased	Living	*	0%
	Deceased	Deceased	Living	Living	*	0%
	Living	Deceased	Deceased	Living	*	0%
ONE Source	Deceased	Living	Living	Living	60	1%
	Living	Living	Deceased	Living	60	1%
	Living	Living	Living	Deceased	40	1%
	Living	Deceased	Living	Living	20	0%
ZERO Sources	Living	Living	Living	Living	4,740	86%

Note: Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Analyses are not weighted. Sample sizes are rounded as per IES Restricted Data security requirements; "*" indicates that the cell rounds to zero.

Table 4. Cross-tabulations of Mortality Status by Data Source

SSDMF					NORC				
Accurint®		Living	Deceased	Total	Accurint®		Living	Deceased	Total
	Living	4,970	30	5,000		Living	4,810	190	5,000
	Deceased	170	300	470		Deceased	70	410	480
	Total	5,140	330	5,470		Total	4,880	600	5,480
		Rate of Agreement		96%			Rate of Agreement		95%
		Kappa		0.74			Kappa		0.73
NDI					NORC				
Accurint®		Living	Deceased	Total	SSDMF		Living	Deceased	Total
	Living	4,790	210	5,000		Living	4,850	290	5,140
	Deceased	80	390	470		Deceased	20	310	330
	Total	4,870	600	5,470		Total	4,870	600	5,470
		Rate of Agreement		95%			Rate of Agreement		94%
		Kappa		0.70			Kappa		0.64
NDI					NDI				
SSDMF		Living	Deceased	Total	NORC		Living	Deceased	Total
	Living	4,840	300	5,140		Living	4,810	60	4,870
	Deceased	40	300	340		Deceased	60	540	600
	Total	4,880	600	5,480		Total	4,870	600	5,470
		Rate of Agreement		94%			Rate of Agreement		98%
		Kappa		0.61			Kappa		0.88

Note : Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Analyses are not weighted. Sample sizes are rounded as per IES Restricted Data security requirements.

Table 5. Percent Deceased by Student Attributes and Data Source

	Accurint®		SSDMF		NDI		NORC	
	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group
<i>Educational Attainment</i>								
Less than High School	11.2%	1.60	8.5%	1.70	14.8%	1.74	14.8%	1.63
High School Graduate	9.5%	1.36	6.5%	1.30	11.7%	1.38	12.1%	1.33
Some College, No BA	8.8%	1.26	6.7%	1.34	11.8%	1.39	11.4%	1.25
BA or Higher [Reference]	7.0%	1.00	5.0%	1.00	8.5%	1.00	9.1%	1.00
<i>Gender</i>								
Male	11.1%	1.71	8.4%	2.00	14.3%	1.74	14.1%	1.66
Female [Reference]	6.5%	1.00	4.2%	1.00	8.2%	1.00	8.5%	1.00
<i>Race/Ethnicity</i>								
White [Reference]	8.9%	1.00	6.6%	1.00	10.8%	1.00	11.2%	1.00
Black	8.3%	0.93	5.8%	0.88	13.5%	1.25	11.6%	1.04
Hispanic	9.5%	1.07	7.1%	1.08	12.6%	1.17	12.8%	1.14
<i>Mother's Educational Attainment</i>								
Less than High School	9.2%	1.32	6.5%	1.13	11.7%	1.06	12.4%	1.12
High School Graduate	10.2%	1.47	7.0%	1.21	11.8%	1.07	11.6%	1.05
Some College or More [Reference]	7.0%	1.00	5.8%	1.00	11.0%	1.00	11.1%	1.00

Note: Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Missing values on student attribute variables imputed using chained equations in Stata's ice command. Analyses weighted by base year sampling weight BYWT. Sample sizes are rounded as per IES Restricted Data security requirements.

Table 6. Logistic Regression Models of Mortality Status, by Data Source

	Accurint®		SSDMF		NDI		NORC	
	Coef.	(se)	Coef.	(se)	Coef.	(se)	Coef.	(se)
Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	-0.15	(0.31)	-0.28	(0.36)	-0.22	(0.25)	-0.21	(0.24)
Some College, No BA	-0.19	(0.26)	-0.20	(0.30)	-0.14	(0.23)	-0.24	(0.23)
BA or Higher	-0.44	(0.35)	-0.58	(0.41)	-0.49	(0.32)	-0.53	(0.30)
Gender								
Male	[Reference]		[Reference]		[Reference]		[Reference]	
Female	-0.55	(0.14) **	-0.70	(0.17) **	-0.60	(0.12) **	-0.59	(0.12) **
Race/Ethnicity								
White	[Reference]		[Reference]		[Reference]		[Reference]	
Hispanic	-0.02	(0.16)	-0.01	(0.20)	0.14	(0.15)	0.09	(0.15)
Black	-0.07	(0.18)	-0.11	(0.22)	0.25	(0.15)	0.07	(0.15)
All Others	0.10	(0.35)	-0.25	(0.36)	0.01	(0.35)	-0.03	(0.33)
Nativity								
Born in United States	[Reference]		[Reference]		[Reference]		[Reference]	
Born Abroad	0.00	(0.27)	0.05	(0.45)	-0.45	(0.30)	-0.26	(0.27)
Father's Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	0.10	(0.19)	0.11	(0.23)	-0.09	(0.17)	0.06	(0.17)
Some College or More	0.28	(0.24)	0.32	(0.29)	0.03	(0.17)	0.19	(0.18)
Mother's Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	0.05	(0.20)	-0.04	(0.23)	0.00	(0.16)	-0.11	(0.18)
Some College or More	-0.41	(0.24)	-0.28	(0.29)	-0.04	(0.19)	-0.15	(0.20)
Academic Achievement in 1980								
Reading Test Score	0.01	(0.01)	0.01	(0.02)	0.03	(0.01) *	0.02	(0.01)
Math Test Score	0.00	(0.01)	0.00	(0.01)	-0.01	(0.01)	-0.01	(0.01)
Self-reported GPA	-0.05	(0.10)	0.00	(0.12)	-0.06	(0.09)	0.03	(0.09)
Non-Cognitive Skills in 1980								
Self-concept	-0.02	(0.07)	-0.04	(0.08)	-0.06	(0.06)	-0.15	(0.07) *
Locus of control	0.01	(0.07)	-0.04	(0.08)	0.03	(0.06)	0.04	(0.07)
Work orientation	0.00	(0.06)	0.08	(0.07)	0.02	(0.06)	-0.01	(0.06)
Health and Disability in 1980								
Body Mass Index	0.01	(0.02)	-0.01	(0.02)	0.01	(0.01)	0.01	(0.01)
Limiting Phys. Condition	0.21	(0.19)	0.31	(0.22)	0.07	(0.18)	0.05	(0.18)
Constant	-2.24	(0.43) **	-2.15	(0.59) **	-1.73	(0.39) **	-1.82	(0.39) **

Note : Sample restricted to the 5,470 HSB sophomore sample members who responded to the 1980 HSB survey but not the 2014 survey. Missing values on predictor variables imputed using chained equations in Stata's ice command. Analyses weighted by base year sampling weight BYWT. Sample sizes are rounded as per IES Restricted Data security requirements. **=p<0.01 *=p<0.05 (two-tailed tests)