How Inferences about Mortality Rates and Gradients Vary by Source of Mortality Information

September 24, 2014

John Robert Warren – University of Minnesota
Carolina Milesi - NORC
Karen Grigorian - NORC
Melissa Humphries – University of Texas-Austin
Eric Grodsky – University of Wisconsin-Madison
Chandra Muller – University of Texas-Austin

Abstract

Researchers who study mortality among participants in longitudinal studies have multiple options for obtaining information about which participants died (and when and how they died). Some use credit bureau-based databases; others use the National Death Index; some use the Social Security Death Index; and still others combine sources and use more genealogical methods. In this paper, we ask how inferences about mortality gradients are altered by the choice of source of mortality information. Using new data on a cohort of 13,479 people who were first interviewed as high school sophomores in 1980 and for whom we have extensive identifying information, we analyze mortality gradients using four separate sources of mortality data. We find that these sources often disagree about which of our panels have died, and also about overall mortality rates. However, our assessments of mortality gradients (i.e., by sex, race/ethnicity, education) are similar regardless of sources of mortality data.

Research on mortality gradients frequently makes use of administrative or other public records on the date and cause of people's deaths. Advances in record linkage technology and computing power, along with new administrative data sharing arrangements, have made possible a new generation of research on the ways in which mortality depends on individuals' social, economic, demographic, and other circumstances while alive. A general presumption in such research is that public administrative data on who has died (and when and how they died) are accurate and reliable.

Investigators who wish to link administrative mortality data to survey or other data now have a number of options for doing so. Many credit bureau-based information brokers (e.g., Accurint™ or Microbilt™)—which are widely used for locating or "tracing" sample members in

preparation for longitudinal follow-up surveys—also report whether and when subjects died. Researchers can pay the National Center for Health Statistics to link records to the National Death Index. The Social Security Death Index can be used by anyone with access to research subjects' names, birth dates, and/or Social Security Numbers. Of course, researchers with access to that identifying information can also use internet search engines and genealogical resources to locate things like obituaries or news reports about people's deaths.

These various sources of information about who has died (and usually when they died) draw on different administrative and public records. Some are based on Social Security Administration records, some are based on credit bureau data, and some draw on states' vital record systems. These various sources thus also have different strengths and weaknesses with respect to population coverage, timeliness of updates, and accuracy. However, researchers who study mortality gradients—that is, who use such mortality information as the basis for their dependent variable in analyses of the socioeconomic, demographic, or other predictors of death—generally assume that their results are not much affected by the source of data on which they base their mortality information.

In this paper, we ask how inferences about mortality gradients are altered by the choice of source of mortality information. Using new data on a cohort of 13,479 people who were first interviewed as high school sophomores in 1980 and for whom we have extensive identifying information, we analyze mortality gradients using four separate sources of mortality data: Accurint™, the Social Security Death Index (SSDI), the National Death Index (NDI), and an approach that begins with these various data resources but then uses targeted web searches and genealogical methods to triangulate and verify information from other data resources.

Because we study the same subjects using the same measures of panel members' social, demographic, and economic attributes, any differences across analyses in inferences about mortality gradients can be attributed to differences across sources of mortality data.

BACKGROUND

[By PAA: Here we will review the literature on mortality gradients, and note that different studies ascertain mortality status and timing differently. Next we will describe the various sources of mortality data in some detail. We will then review what we already know about differences in coverage and quality of mortality records. We will also review efforts to compare inferences about who has died across sources of information. Finally, we will review the few previous efforts to assess whether inferences about mortality gradients depend on data source]

RESEARCH DESIGN

[Here we will describe the HS&B sophomore study. Then we will say exactly how we linked to the several sources of mortality data. Finally, we will describe how we (a) constructed the social, economic, demographic, and academic predictor variables and (b) imputed missing values using chained equations. The latter description will refer to Table 1, which includes descriptive statistics on each variable before and after imputation.]

[Here, “Accurint” means both the 2012 and 2013 sweeps; anyone identified as dead on either are classified as dead. “SSDI-SSN” means that a sample member was identified as dead because their SSN exactly matched a record in the SSDI. “SSDI-Name 1” starts with SSDI-SSN but then also classifies as dead anyone whose exact date of birth and name match a record in the SSDI. “SSDI-Name 2” starts with SSDI-Name 1 but then also classifies as dead any women whose exact date of birth and first name match a record in the SSDI. Finally, “NORC” means that the sample member’s final disposition code is “deceased.” We will add NDI later.]

RESULTS

Table 2 describes the number and percentage dead among the 13,749 sophomore sample members, separately by source of mortality data. Accurint™, SSDI-Name 1, and NORC produce similar mortality rates, with between 4.0% and 4.02% deceased. SSDI-SSN produced a much lower mortality rate (3.0%) and SSDI-Name 2 produced a much higher rate (16.5%); the latter two are not surprising given the high likelihood of false negatives and false positives from those sources.

The fact that some of the data sources produce similar mortality rates does not necessarily mean that those data sources agree about exactly which sample members died. In Table 3 we cross-classify mortality status (living vs. deceased) for each pair of sources of information about mortality. Sample members for whom sources of information disagree about mortality status appear in the off-diagonal cells of each cross-classification table.

Excluding SSDI-Name 2 (which tends to produce large numbers of false positives), between 1.8% and 3.8% of sample members are in the off-diagonal cells. Although these disagreement rates are low, this still means that hundreds of sample members are classified as living or deceased differently by different data sources. Also, this overall rate of agreement combined with generally low mortality rates masks the fact that large percentages of classified as deceased by one data resource are classified as living by others.

Even disagreement across data sources about which sample members are deceased does not necessarily mean that inferences about mortality gradients will be affected by choice of information source about mortality status. In Table 4, we report sample members' mortality status separately by several of the individual-level attributes described in Table 1: educational attainment, gender, race/ethnicity, and mother's educational attainment (as a proxy for family socioeconomic background). [Later, we will add figures that depict these statistics and that add error bars; that way we can formally assess whether differences are significant.] Generally speaking, descriptions about mortality differentials depend on which source of information we use to classified sample members' mortality status.

How is educational attainment associated with mortality? Based on data from Accurint™ or SSDI-SSN, high school dropouts are about 3 times as likely as college graduates to be deceased. Based on NORC's final disposition codes, however, they are 4 times as likely to be deceased. Even differentials between college graduate and those who attend college but do not complete BAs are larger using information from NORC.

How does mortality status depend on gender or race/ethnicity? Men are about twice as likely to have died, unless we use SSDI-Name 1 (where they are only 1.5 times as likely to

have died). Across data sources, African Americans are between 43% and 63% more likely to have died; Hispanic-White differentials are more consistent, with Hispanics between 46% and 58% more likely to be deceased.

Finally, how does risk of mortality vary across socioeconomic background (as evidenced by mothers' levels of completed schooling)? Compared to children whose mothers attended college, those whose mothers dropped out of high school were anywhere between 32% and 74% more likely to have died. That is, our inference about the bivariate relationship between family socioeconomic background and mortality would differ greatly depending on how we ascertained which sample members are dead.

Serious analyses of mortality gradients would probably use multivariate regression techniques to understand the ways in which mortality status and timing of death vary by socioeconomic, demographic, and other circumstances. In Table 5 we report results of logistic regression models in which mortality status (0 equals living, 1 means deceased) is a function of all of the predictors described in Table 1. We estimate separate models using mortality information from Accurint™, SSDI-SSN, SSDI-Name 1, and NORC.

Women are always significantly less likely to have died, although the point estimate describing the magnitude of that conditional association varies. All else constant, there are no significant race/ethnic differentials (except that African Americans are more likely to have died according to SSDI-Name 1). There are no significant conditional associations between mortality status and sample members' non-cognitive skills or early life health.

In contrast, inferences about the statistical significance of mortality differentials by educational attainment, mother's educational attainment, and cognitive skills depend on how

we ascertain mortality status. For example, reading achievement test scores and grade point average are independently and significantly associated with mortality, but only when we base mortality on SSDI-Name 1. Mothers' education is independently and significantly associated with mortality, but only using mortality data from Accurint™. The net association between completing college and mortality is statistically significant, except when we use SSDI-SSN.

However, even though the statistical significance of certain point estimates may vary across sources of mortality data, it is not the case that the coefficients significantly differ from one another across models. [Later, we will add figures that depict these coefficients and that add error bars; that way we can formally assess whether differences are significant. They are not, though.] For example, 95% confidence intervals for the coefficient describing the conditional association between gender and mortality overlap; the same is true for 95% confidence intervals for the coefficient for college completion. In other words, the source of mortality data may influence our inferences about whether particular coefficients are significantly different from zero; however, our assessment of the magnitude of those coefficients is not necessarily influenced by choice of data source.

DISCUSSION

[To be completed later, once the results are finalized]

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

	Prior to Imputation					After Imputation (n=13,749)			
	Mean / %	(sd)	n	Min.	Max.	Mean / %	SD	Min	Max
Educational Attainment									
Less than High School	4.5%		12,131	0.0	1.0	5.2%		0.0	1.0
High School Graduate	30.2%		12,131	0.0	1.0	31.0%		0.0	1.0
Some College, No BA	40.2%		12,131	0.0	1.0	39.9%		0.0	1.0
BA or Higher	25.2%		12,131	0.0	1.0	24.0%		0.0	1.0
Gender									
Male	49.7%		13,749	0.0	1.0	49.7%		0.0	1.0
Female	50.3%		13,749	0.0	1.0	50.3%		0.0	1.0
Race/Ethnicity									
White	72.4%		13,749	0.0	1.0	72.4%		0.0	1.0
Hispanic	12.8%		13,749	0.0	1.0	12.8%		0.0	1.0
Black	12.1%		13,749	0.0	1.0	12.1%		0.0	1.0
All Others	2.6%		13,749	0.0	1.0	2.6%		0.0	1.0
Nativity									
Born in United States	95.7%		13,529	0.0	1.0	95.6%		0.0	1.0
Born Abroad	4.3%		13,529	0.0	1.0	4.4%		0.0	1.0
Father's Educational Attainment									
Less than High School	23.7%		9,448	0.0	1.0	26.0%		0.0	1.0
High School Graduate	32.2%		9,448	0.0	1.0	32.1%		0.0	1.0
Some College or More	44.1%		9,448	0.0	1.0	41.9%		0.0	1.0
Mother's Educational Attainment									
Less than High School	19.3%		10,825	0.0	1.0	20.6%		0.0	1.0
High School Graduate	44.9%		10,825	0.0	1.0	44.5%		0.0	1.0
Some College or More	35.8%		10,825	0.0	1.0	34.9%		0.0	1.0
Academic Achievement in 1980									
Reading Test Score	6.8	(4.8)	12,549	-1.3	19.0	7.8	(5.7)	-1.3	19.0
Math Test Score	12.4	(9.8)	12,549	-4.5	38.0	12.9	(10.5)	-4.5	38.0
Self-reported GPA	2.7	(0.8)	13,609	0.0	4.0	2.7	(0.8)	0.0	4.0
Non-Cognitive Skills in 1980									
Self-concept	0.0	(1.0)	13,153	-4.8	1.7	0.0	(1.0)	-4.8	1.7
Locus of control	0.0	(1.0)	13,130	-4.2	2.2	0.0	(1.0)	-4.2	2.2
Work orientation	0.0	(1.0)	13,255	-6.3	1.7	0.0	(1.0)	-6.3	1.7
Health and Disability in 1980									
Body Mass Index	21.1	(3.3)	12,539	7.8	122.6	21.1	(3.3)	7.8	122.6
Limiting Physical Condition	8.3%		12,744	0.0	1.0	8.5%		0.0	1.0

Note : Sample restricted to High School & Beyond (HS&B) sophomore sample members who responded to the 1980 HS&B survey. Analyses weighted by base year sampling weight BYWT.

Table 2. Descriptive Statistics for Mortality Outcomes

	Living	Deceased	% Deceased
Accurint®	13,204	545	4.0%
SSDI-SSN: Social Security Number Matches Exactly	13,448	301	2.2%
SSDI-Name 1: Social Security Number Matches Exactly —or — (Exact Date of Birth Matches —and— Name Matches)	13,181	568	4.1%
SSDI-Name 2: Social Security Number Matches Exactly —or — (Exact Date of Birth Matches —and— Name Matches) — or — (Exact Date of Birth Matches —and— Woman's First Name Matches)	11,480	2,269	16.5%
NORC: Final Disposition After 2013 Locating Effort	13,168	581	4.2%

Note : SSDI=Social Security Death Index. Sample restricted to the 13,749 High School & Beyond (HS&B) sophomore sample members who responded to the 1980 HS&B survey. Analyses are unweighted.

Table 3. Cross-tabulations of Mortality Status by Data Source

Accurint®	SSDI-SSN: Living	SSDI-SSN: Deceased	*Total*
Living	13,203	1	13,204
Deceased	245	300	545
Total	13,448	301	13,749
Percentage in Off-Diagonal Cell:			1.8%

SSDI-SSN	SSDI-Name 2: Living	SSDI-Name 2: Deceased	*Total*
Living	11,480	1,968	13,448
Deceased	0	301	301
Total	11,480	2,269	13,749
Percentage in Off-Diagonal Cell:			14.3%

Accurint®	SSDI-Name 1: Living	SSDI-Name 1: Deceased	*Total*
Living	12,975	229	13,204
Deceased	206	339	545
Total	13,181	568	13,749
Percentage in Off-Diagonal Cell:			3.2%

SSDI-SSN	NORC: Living	NORC: Deceased	*Total*
Living	13,138	310	13,448
Deceased	30	271	301
Total	13,168	581	13,749
Percentage in Off-Diagonal Cell:			2.5%

Accurint®	SSDI-Name 2: Living	SSDI-Name 2: Deceased	*Total*
Living	11,294	1,910	13,204
Deceased	186	359	545
Total	11,480	2,269	13,749
Percentage in Off-Diagonal Cell:			15.2%

SSDI-Name 1	SSDI-Name 2: Living	SSDI-Name 2: Deceased	*Total*
Living	11,480	1,701	13,181
Deceased	0	568	568
Total	11,480	2,269	13,749
Percentage in Off-Diagonal Cell:			12.4%

Accurint®	NORC: Living	NORC: Deceased	*Total*
Living	13,027	177	13,204
Deceased	141	404	545
Total	13,168	581	13,749
Percentage in Off-Diagonal Cell:			2.3%

SSDI-Name 1	NORC: Living	NORC: Deceased	*Total*
Living	12,913	268	13,181
Deceased	255	313	568
Total	13,168	581	13,749
Percentage in Off-Diagonal Cell:			3.8%

SSDI-SSN	SSDI-Name 1: Living	SSDI-Name 1: Deceased	*Total*
Living	13,181	267	13,448
Deceased	0	301	301
Total	13,181	568	13,749
Percentage in Off-Diagonal Cell:			1.9%

SSDI-Name 2	NORC: Living	NORC: Deceased	*Total*
Living	11,244	236	11,480
Deceased	1,924	345	2,269
Total	13,168	581	13,749
Percentage in Off-Diagonal Cell:			15.7%

Note : See Table 2 for defitintion of mortality data sources. Sample restricted to High School & Beyond (HS&B) sophomore sample members who responded to the 1980 HS&B survey. Analyses are unweighted.

Table 4. Percent Deceased by Student Attributes and Data Source

	Accurint®		SSDI-SSN		SSDI-Name 1		NORC	
	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group	% Deceased	Rate vs. Reference Group
Educational Attainment								
Less than High School	7.4%	3.08	4.3%	3.07	8.0%	2.67	8.7%	4.14
High School Graduate	4.7%	1.96	2.4%	1.71	4.6%	1.53	5.4%	2.57
Some College, No BA	3.8%	1.58	2.2%	1.57	4.6%	1.53	4.3%	2.05
BA or Higher **[Reference]**	2.4%	1.00	1.4%	1.00	3.0%	1.00	2.1%	1.00
Gender								
Male	5.4%	2.08	3.0%	2.14	5.3%	1.51	5.8%	2.00
Female **[Reference]**	2.6%	1.00	1.4%	1.00	3.5%	1.00	2.9%	1.00
Race/Ethnicity								
White **[Reference]**	3.5%	1.00	1.9%	1.00	3.9%	1.00	3.8%	1.00
Black	5.0%	1.43	3.1%	1.63	6.3%	1.62	5.6%	1.47
Hispanic	5.1%	1.46	2.8%	1.47	5.7%	1.46	6.0%	1.58
Mother's Educational Attainment								
Less than High School	4.7%	1.74	2.4%	1.41	4.9%	1.32	5.6%	1.56
High School Graduate	4.7%	1.74	2.5%	1.47	4.7%	1.27	4.4%	1.22
Some College or More **[Reference]**	2.7%	1.00	1.7%	1.00	3.7%	1.00	3.6%	1.00

Note : See Table 2 for defitintion of mortality data sources. Sample restricted to the 13,749 High School & Beyond (HS&B) sophomore sample members who responded to the 1980 HS&B survey. Analyses weighted by base year sampling weight BYWT.

Table 5. Logistic Regression Models of Mortality Status, by Data Source

	Accurint®		SSDI-SSN		SSDI-Name 1		NORC	
	Coef.	(se)	Coef.	(se)	Coef.	(se)	Coef.	(se)
Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	-0.35	(0.29)	-0.49	(0.41)	-0.47	(0.27)	-0.38	(0.26)
Some College, No BA	-0.43	(0.25)	-0.48	(0.35)	-0.37	(0.24)	-0.49	(0.23) *
BA or Higher	-0.75	(0.31) *	-0.87	(0.48)	-0.68	(0.29) *	-1.11	(0.31) **
Gender								
Male	[Reference]		[Reference]		[Reference]		[Reference]	
Female	-0.71	(0.13) **	-0.68	(0.17) **	-0.39	(0.12) **	-0.70	(0.12) **
Race/Ethnicity								
White	[Reference]		[Reference]		[Reference]		[Reference]	
Hispanic	0.17	(0.15)	0.19	(0.19)	0.26	(0.15)	0.24	(0.15)
Black	0.27	(0.16)	0.41	(0.21)	0.45	(0.15) **	0.28	(0.16)
All Others	0.35	(0.32)	0.00	(0.36)	-0.38	(0.31)	0.23	(0.33)
Nativity								
Born in United States	[Reference]		[Reference]		[Reference]		[Reference]	
Born Abroad	0.14	(0.26)	0.23	(0.32)	0.33	(0.30)	-0.01	(0.27)
Father's Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	0.09	(0.19)	0.07	(0.23)	-0.09	(0.18)	-0.01	(0.18)
Some College or More	0.30	(0.23)	0.30	(0.31)	-0.12	(0.20)	0.17	(0.19)
Mother's Educational Attainment								
Less than High School	[Reference]		[Reference]		[Reference]		[Reference]	
High School Graduate	-0.02	(0.18)	0.03	(0.24)	0.05	(0.16)	-0.15	(0.18)
Some College or More	-0.53	(0.22) *	-0.29	(0.29)	-0.08	(0.19)	-0.23	(0.21)
Academic Achievement in 1980								
Reading Test Score	0.01	(0.01)	0.02	(0.02)	0.03	(0.01) *	0.02	(0.01)
Math Test Score	-0.01	(0.01)	-0.01	(0.01)	0.00	(0.01)	-0.01	(0.01)
Self-reported GPA	-0.11	(0.09)	-0.15	(0.13)	-0.17	(0.09) *	-0.07	(0.09)
Non-Cognitive Skills in 1980								
Self-concept	-0.01	(0.06)	0.00	(0.08)	-0.02	(0.06)	-0.12	(0.08)
Locus of control	0.00	(0.06)	0.05	(0.08)	-0.02	(0.06)	-0.01	(0.07)
Work orientation	-0.02	(0.06)	0.05	(0.07)	0.06	(0.06)	0.00	(0.06)
Health and Disability in 1980								
Body Mass Index	0.03	(0.01)	0.00	(0.02)	-0.03	(0.02)	0.01	(0.01)
Limiting Physical Conditio	0.18	(0.18)	0.39	(0.21)	-0.02	(0.18)	0.14	(0.18)
Constant	-2.84	(0.41) **	-2.82	(0.62) **	-1.58	(0.52) **	-2.39	(0.38) **

Note : See Table 2 for defitintion of mortality data sources. Sample restricted to the 13,749 High School & Beyond (HS&B) sophomore sample members who responded to the 1980 HS&B survey. Analyses weighted by base year sampling weight BYWT. **=p<0.01 *=p<0.05 (two-tailed tests)