Am J Cancer Res 2013;3(5):518-522
www.ajcr.us /ISSN:2156-6976/ajcr0000225

Original Article

The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients

Justin T Huntington[1], Mathew Butterfield[1], James Fisher[2], Daniel Torrent[3], Mark Bloomston[1]

[1]*Division of Surgical Oncology, The Ohio State University Wexner Medical Center, Columbus, OH, USA;* [2]*College of Public Health, The Ohio State University Wexner Medical Center, Columbus, OH, USA;* [3]*Department of Surgery, East Carolina University, Greenville, NC, USA*

Received August 28, 2013; Accepted September 29, 2013; Epub November 1, 2013; Published November 15, 2013

Abstract: Introduction: The ability to ascertain survival information is important for retrospective and prospective studies. Two databases that can be used are the Social Security Death Index (SSDI) and the National Death Index (NDI). Although the NDI is more complete, there are advantages to the SSDI such as ease of use and cost. The intent of this study was to determine accuracy of the SSDI. Methods: Publically available data on all known deceased individuals in the state of Ohio in 2003 were obtained from the State of Ohio Department of Health. A random sample of 63,557 of these were compared to the SSDI to identify risk factor for inclusion/exclusion. Results: Overall, 94.7% of all death records were confirmed by the SSDI. Age at death, gender, race, ethnicity, and cause of death were all found to significantly affect the likelihood of inclusion. Specifically, people aged 18-24 were included only 79.8% of the time compared to 96.2% for those over the age of 65. Also, malignancy as cause of death resulted in a 95.3% inclusion while trauma as a cause of death led to 86.5% inclusion. While Caucasians had an inclusion of 95.6%, African Americans were included only 87.8% of the time. Hispanics and women also had lower inclusion rates. Discussion: The SSDI is a strong tool for following up on participants lost to follow up in certain populations but is weaker in others. The SSDI would be particularly useful in a population that is largely older, Caucasian, or has malignant disease.

Keywords: Social, security, death, index, SSDI, NDI, survival

Introduction

Obtaining accurate follow-up is often difficult but of the utmost importance for retrospective and prospective studies. Survival assessment if of particular importance in cancer-related studies but may be influenced by study participants who are lost to follow-up, thus leading to bias if there are systematic differences between the group that is lost and those that remained [1]. Therefore, the ability to identify study patients who have died becomes and important step in maintaining study databases and registries where survival is important. The Social Security Administration maintains a death master file, which may be publically searched, online as the Social Security Death Index (SSDI, http://ssdi.rootsweb.ancestry.com). This database is commonly utilized owing to its ease of use to ascertain death for research purposes [2, 3]. To date, however, no large-scale study has documented its accuracy or delineated a population of patients for which it is best suited.

The National Death Index (NDI), created in 1979, has long been considered the "gold standard" among national searchable databases and contains greater than 95% of deaths from individual state records from 1979 through the present [2, 4]. However, the NDI is not readily accessible and is expensive and slow to use. Conversely, the SSDI can be searched online (http://ssdi.rootsweb.ancestry.com) and is free of cost [2-5]. Also, additions to the NDI are done annually which results in a twelve to twenty-four month delay in a new death appearing on the database [6]. The SSDI, on the other hand, is frequently updated [4, 7]. While other small studies have reported the accuracy of SSDI using the NDI as the gold standard [2-11], the purpose of this study was to verify the accuracy

Table 1. Proportion of Ohio death records included in Social Security Death Index

	Included	Excluded	Total	Inclusion	p
Age					< 0.0001
18-24	435	110	545	79.8%	
25-34	714	152	866	82.4%	
35-44	1,669	327	1,996	83.6%	
45-54	3,766	466	4,232	89%	
55-64	6,401	442	6,843	93.5%	
65+	47,224	1,851	49,075	96.2%	
Gender					0.028
Male	28,598	1,525	30,123	94.9%	
Female	31,611	1,823	33,434	94.5%	
Cause of Death					< 0.0001
Malignant	14,159	699	14,858	95.3%	
Trauma	2,513	393	2,906	86.5%	
Other	43,187	2,245	45,432	95.1%	
Race					< 0.0001
White	54,211	2,518	56,729	95.6%	
Black	5,837	812	6,649	87.8%	
Other	161	18	179	89.9%	
Hispanic					< 0.0001
Yes	313	39	352	88.9%	
No	59,793	3,203	62,996	94.9%	
Total	60,209	3,348	63,557	94.7%	

of the SSDI using publically available state death records and to identify potential risk factors for exclusion.

Methods

Public death records of Ohio residents in the year 2003 were provided free of charge by the Ohio Department of Health Center for Vital and Health Statistics. Data provided included date of birth, date of death, age at death, gender, cause of death, marital status, residence within city limits, history of military service, race, and Hispanic ethnicity. Records were initially excluded if the person was less than 18 years old at time of death or if there was no social security number. Others were later excluded if the social security number provided did not match the person in the record when searched against the SSDI. For records where only the surname was different after checking against the SSDI, identity was verified by matching dates of birth and death between the Ohio records and the SSDI.

A random sample of 63,557 social security numbers from the Ohio death records was entered into the SSDI. If a record was returned which matched that of the Ohio record, then this record was considered included in the SSDI. If no match was found, the social security number was reentered to account for the possibility of entering the number incorrectly. If there was still no match this was considered non-inclusion in the SSDI. If the wrong record was returned, this was considered a wrong social security number in the Ohio Department of Health Center for Vital and Health Statistics record and this record was excluded from the study as per the exclusion criteria. If the same first name but different surname was returned then identity was matched by matching birth and death dates between the two sources.

Statistical analysis

Age was coded as a continuous variable in the original data but was converted to 6 categorical variables for statistical analysis. Cause of death was based on the National Center for Health Statistics list of 50 leading causes of death but was converted into a three level categorical variable of malignancy, trauma, and other (reference). Malignancy was included in the original list of 50 while trauma was created by combining unintentional injuries, intentional self-harm (suicide), assault (homicide), and legal intervention.

The crude proportion included in the SSDI from the groups in each variable was assessed first. Differences in dichotomous variables (e.g. gender, Hispanic) were compared using a two-group proportion test. For variables with more than two groups (e.g. age, race, cause of death), ANOVA was used to look for differences. For significant ANOVA results, Scheffe test was used to see which comparisons were significant at an alpha level of 0.05.

The logistic regression model was built by including all variables. Hierarchical backwards selection was run on this model. Likelihood ratio chi-square was used to obtain *p*-values,

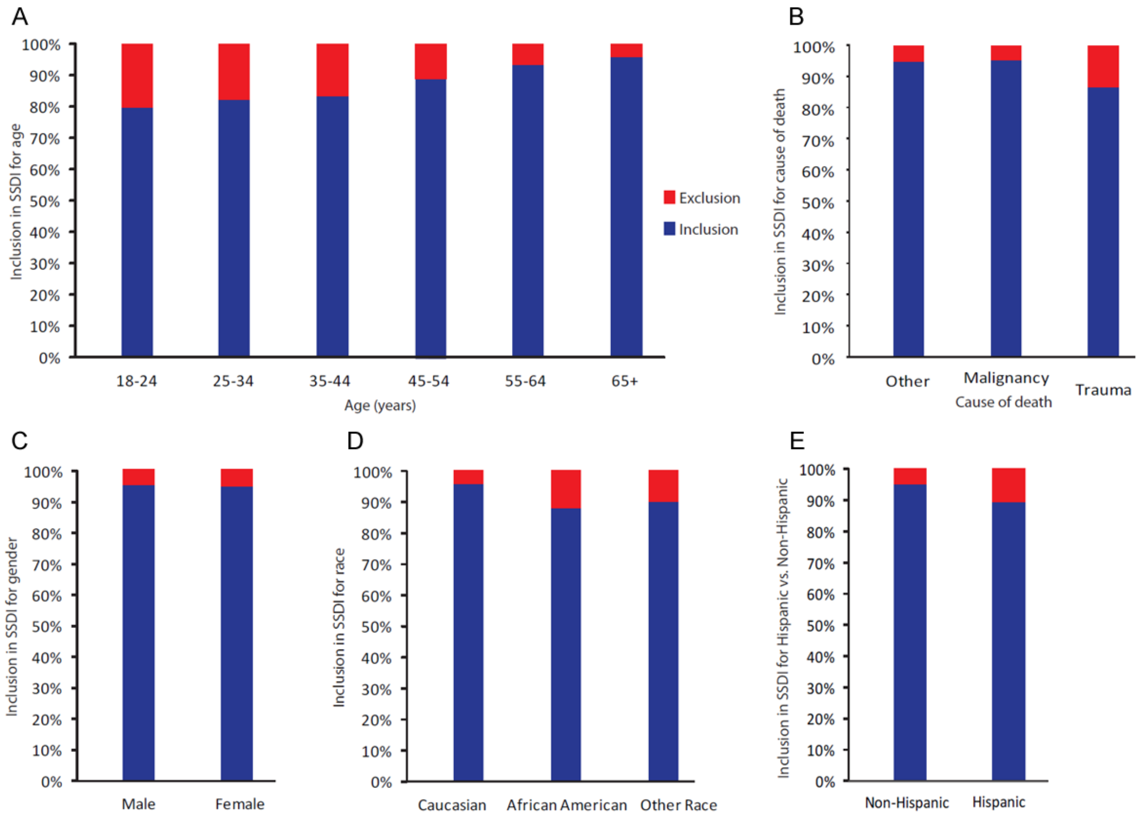

Figure 1. The overall inclusion in the SSDI by category. The overall percentage included in the SSDI is shown here. A: Inclusion by age in 6 categories. B: Inclusion by cause of death (malignancy, trauma, other). C: Inclusion by gender. D: Inclusion by race (Caucasian, African American, other race). E: Inclusion for Hispanic versus non-Hispanic persons.

which were compared to an alpha of 0.05 for inclusion in the model. Four interaction terms were included in the initial model (Hispanic and cause of death, Hispanic and age, race and cause of death, race and age), but none were significant. For the remaining variables, place of death was not significant and was removed from the model. After being removed it was added back to assess for confounding and was judged to not significantly affect the odds ratios.

Goodness of fit of the final model was assessed using the Hosmer-Lemeshow test with a cut off alpha of 0.05. A non-significant *p*-value indicated that this model fit the data well. All analyses were undertaken using Stata (StataCorp LP, College Station, TX).

Results

There were 63,557 records evaluated and overall 94.7% were included in the SSDI (**Table 1** and **Figure 1**). The percent included in the SSDI differed by age group as listed by category ($p < 0.0001$ with an alpha level of 0.5). As age increased, so did the likelihood of being included in the SSDI with the highest age group (65+) being most likely to be included (96.2%). As expected, this oldest category accounted for the vast majority of deaths (77%). Men were more likely to be included ($p = 0.028$) as were those dying from malignancy ($p < 0.0001$). Those dying from trauma were the least likely to be included. There were also differences in inclusion by race ($p < 0.0001$) with the inclusion for Caucasians being higher than inclusion for African Americans or other at an alpha level of 0.05. The difference between African Americans and other was not significant. The difference in inclusion between Hispanics and non-Hispanics was significant ($p < 0.0001$).

Multivariate logistic regression analysis was then undertaken utilizing all measured values from the 63,040 sets of records that had complete data to determine which of these factors

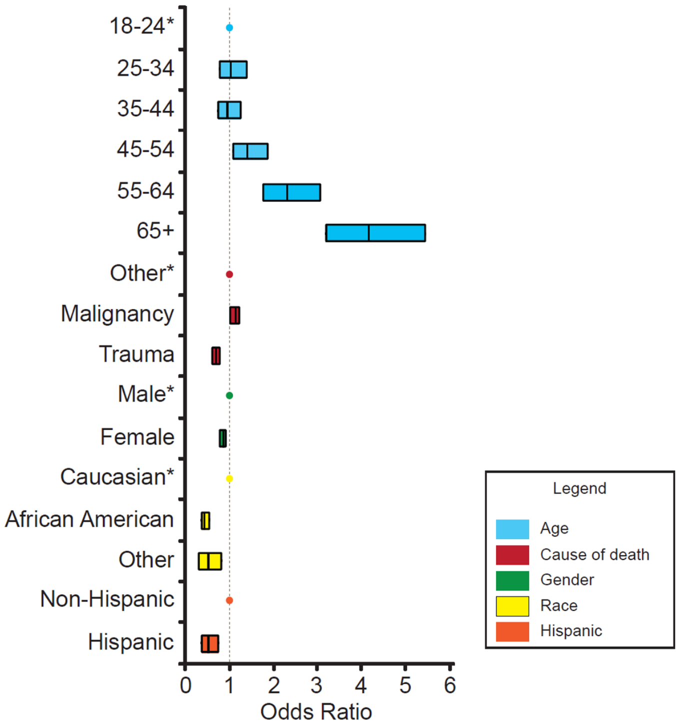

Figure 2. Multivariate logistic regression analysis of inclusion in the SSDI. An asterisk is shown next to each variable that was used as the referent value for comparison for the other variables in that category. The dotted line represents an odds ratio of 1. Anything to the right of this line represents favorable inclusion when compared to the referent variable whereas those to the left are less likely to be included in the SSDI when compared to the variant. The line within each bar represents the odds ratio whereas the bars themselves represent the 95% confidence interval for each variable. The categories that were examined were age, cause of death, gender, race, and Hispanic versus non-Hispanic.

would significantly impact inclusion in the SSDI. Only older age and death from malignancy significantly increased the likelihood of being included (**Figure 2**). Age ranges 45-54 (OR 1.4, p = 0.012), 55-64 (OR 2.3, p < 0.001), and 65+ (OR 4.1, p < 0.001) were more likely to be included than younger adults. Those who died from malignancy were more likely to be included in the SSDI compared to those who died of other causes (OR 1.1, p = 0.047). Those who died of trauma were less likely to be included than those who died from other causes (OR 0.68, p < 0.001). Women were also less likely to be included in the SSDI (OR 0.83, p < 0.0001) adjusting for other variables. Other factors predictive of *not* being included in the SSDI were non-Caucasian race (African American OR 0.41, p < 0.001; "other" OR 0.49, p = 0.005) and Hispanic ethnicity (OR 0.51, p = 0.0006).

Discussion

In this study, we set out to determine the accuracy of the Social Security Death Index. This is the first study of its kind to do so on a large scale. As well, we are the first to utilize Department of Health death records as the gold standard. Overall, we found inclusion in the SSDI to be 95%. In unadjusted analysis, we found that demographic factors and cause of death influence inclusion in the SSDI. Older age, Caucasian race, and male gender were associated with increased probability of inclusion. Being Hispanic and death from trauma were associated with lower probability of inclusion.

We utilized over 65,000 records, which represents more than half of the nearly 110,000 deaths in 2003 in the state of Ohio. Logistic regression was used to adjust for the variables that may affect inclusion. Although some records were not used because of missing data these were less than 1% of the total sample size and likely had a negligible effect on the final outcome. The differences found in the unadjusted analysis held up in the logistic model. The final model included marital status, previous military service, and residence within city limits in addition to the variables of interest.

One potential weakness in the study was that the death records searched against the SSDI were all provided by Ohio. The state does have geographic and racial heterogeneity that should mitigate this limitation yet there may be some regional bias that is unable to be accounted for. A second limitation is the time between the year of death and the year of the SSDI search. We chose a remote year to allow adequate time for deaths to be reported for inclusion in SSDI. Records are added to the SSDI fairly frequently and it's unlikely that many of those found were added more than seven years after death; however, future studies might benefit from using a shorter interval between year of death and year of search.

Based on the results of this study it appears that the SSDI would be a useful tool for ascertaining death in certain studies and weaker in others. It would not be advisable to use the SSDI as the main form of follow up in a study following a population that is primarily younger, composed of mostly racial minorities, or in trauma patients. Studies using the SSDI as the primary means of follow-up should be reviewed cautiously in these patient populations. In older patients it would be reasonable to use the SSDI as a primary source in determining whether a patient who was lost to follow-up has died. This is especially true in an elderly (65+) population where inclusion is greater than 95%. Patients who died from malignant disease also had an unadjusted probability of inclusion of greater than 95% and it would be reasonable to use the SSDI as a main source for determining if those lost to follow up have died in this population. The category for other causes of death was also over 95% but whether that applies to all the individual different causes of death within that group was beyond the scope of this study.

Finally, this study still begs the question of whether somebody not listed in the SSDI can safely be assumed to be still alive. In essence, it is easier to verify somebody had died rather than to prove they are still alive. Reasons why somebody who has died may not be in SSDI include failure to report the death to the Social Security Administration, lack of participation in the social security program, survivor death benefits being paid to dependents or spouse, or human error. These issues not withstanding, SSDI still appears to be a reliable rapid source for providing survival information.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Mark Bloomston, Division of Surgical Oncology, The Ohio State University Wexner Medical Center, N924 Doan Hall, 410 W. 10th Ave, Columbus, OH 43210, USA. Tel: 614-293-4583; Fax: 614-293-4583; E-mail: Mark.Bloomston@osumc.edu

References

[1] Tripepi G, Jager KJ, Dekker FW, Zoccali C. Bias in clinical research. Kidney Int 2008; 73: 148-53.

[2] Sesso HD, Paffenbarger RS, Lee IM. Comparison of national death index and world wide web death searches. Am J Epidemiol 2000; 152: 107-11.

[3] Wojcik NC, Huebner WW, Jorgensen G. Strategies for Using the National Death Index and the Social Security Administration for Death Ascertainment in Large Occupational Cohort Mortality Studies. Am J Epidemiol 2010; 172: 469-477.

[4] Kraut A, Chan E, Landrigan PJ. The costs of searching for deaths: national death index vs social security administration. Am J Public Health 1992; 82: 760-1.

[5] Hermansen SW, Leitzmann MF, Schatzkin A. The impact on national death index ascertainment of limiting submissions to social security administration death master file matches in epidemiologic studies of mortality. Am J Epidemiol 2008; 169: 901-8.

[6] Sohn MW, Arnold N, Maynard C, Hynes DM. Accuracy and completeness of mortality data in the Department of Veterans Affairs. Popul Health Met 2006; 4: 2.

[7] Buchanich JM, Dolan DG, Marsh GM, Madrigano J. Underascertainment of Deaths using Social Security Records: A Recommended Solution to a Little-Known Problem. Am J Epidemiol 2005; 162: 193-194.

[8] Quinn J, Kramer N, McDermott D. Validation of the Social Security Death Index (SSDI): An Important Readily-Available Outcomes Database for Researchers. West J Emerg Med 2007; 9: 6-8.

[9] Schisterman EF, Whitcomb BW. Use of the social security administration death master file for ascertainment of mortality status. Popul Health Met 2004; 2: 2.

[10] Wentworth DN, Neaton JD, Rasmussen WL. An evaluation of the social security administration master beneficiary record fie and the national death index in the ascertainment of vital status. Am J Public Health 1983; 73: 1270-4.

[11] Lash LL, Silliman RA. A Comparison of the National Death Index and Social Security Administration Databases to Ascertain Vital Status. Epidemiology 2001; 12: 259-261.