Original Report:
Big Data
and its Application
in Health Research

# Racial and Ethnic Differences in a Linkage with the National Death Index

Eric A. Miller, PhD, MSPH[1]; Frances A. McCarty, PhD[1]; Jennifer D. Parker, PhD[1]

**Objectives:** Differences in the availability of a Social Security Number (SSN) by race/ethnicity could affect the ability to link with death certificate data in passive follow-up studies and possibly bias mortality disparities reported with linked data. Using 1989-2009 National Health Interview Survey (NHIS) data linked with the National Death Index (NDI) through 2011, we compared the availability of a SSN by race/ethnicity, estimated the percent of links likely missed due to lack of SSNs, and assessed if these estimated missed links affect race/ethnicity disparities reported in the NHIS-linked mortality data.

**Methods:** We used preventive fraction methods based on race/ethnicity-specific Cox proportional hazards models of the relationship between availability of SSN and mortality based on observed links, adjusted for survey year, sex, age, respondent-rated health, education, and US nativity.

**Results:** Availability of a SSN and observed percent linked were significantly lower for Hispanic and Asian/Pacific Islander (PI) participants compared with White non-Hispanic participants. We estimated that more than 18% of expected links were missed due to lack of SSNs among Hispanic and Asian/PI participants compared with about 10% among White non-Hispanic participants. However, correcting the observed links for expected missed links appeared to only have a modest impact on mortality disparities by race/ethnicity.

**Conclusions:** Researchers conducting analyses of mortality disparities using the NDI or other linked death records, need to be cognizant of the potential for differential linkage to contribute to their results. *Ethn Dis*. 2017;27(2):77-84; doi:10.18865/ed.27.2.77

## Introduction

Passive surveillance of vital status is a commonly used method to examine determinants of mortality among race and ethnic subgroups in public health studies.[1-5] In this method, persons in a defined study population are linked with administrative data from death certificates using personally identifiable information (PII) in each data source. Commonly used PII includes full name, date of birth, and Social Security Number (SSN), if available. As a unique identifier, SSN is extremely valuable in data linkage and has been shown to greatly increase sensitivity in linkages with mortality databases.[6] Accurate comparisons of mortality among groups using passive surveillance requires an unbiased ability to link records from the different groups. However, differences in the availability of a SSN, as well as certain population characteristics within race and ethnic groups, can make some of these groups more difficult to link than others. Consequently, differences in the ability to link (differential linkage) by race and ethnicity can potentially produce artificial disparities or bias mortality disparities. In addition, differential linkage can lead to mis-specified determinants of mortality if factors related to linkage are related to mortality.

There is evidence that differential linkage contributes, in part, to the phenomenon known as "The Hispanic Paradox," in which the Hispanic population experiences lower mortality rates than the White non-Hispanic population despite having a higher risk profile.[7,8] For example, using National Health Interview Survey (NHIS) data, Lariscy (2011) found that Hispanic survey participants who linked with death certificate records in the National Death Index (NDI) matched on fewer fields compared with non-Hispanic Whites and the difference was even greater for Hispanic participants born outside the US.[9] The Hispanic population can be

**Keywords:** Record Linkage; Racial Disparities; Social Security Number

[1]National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland, United States

Address correspondence to: Eric A. Miller; National Cancer Institute; 9609 Medical Center Drive; Rockville, MD 20850; 240.276.5336; EricMiller2@nih.gov

more difficult to link than the Black and White non-Hispanic population because of different naming conventions that the linkage process is not necessarily equipped to handle and they have been found to be less likely to have a SSN, especially among an older population and those born outside the US.[9-11] Although there are other factors that contribute to lower mortality rates in the Hispanic population,[12] differential linkage is not commonly acknowledged as a potential contributor, or even an explanation, in studies relying on passive surveillance.

> *…differences in the availability of a SSN, as well as certain population characteristics within race and ethnic groups, can make some of these groups more difficult to link than others.*

For reasons similar to those in the Hispanic population, difficulties with linkage seem likely to exist in the Asian population. Few studies examining bias in record linkage have included the Asian population.[13] The paucity of literature is likely due to few data sources with sufficient sample size to examine this issue. Because the Asian population is rapidly growing in the US,[14] efforts are needed to better understand the health of this population and inaccurate measurements of mortality from passive surveillance can undermine this effort.

The NDI is an important source of passive follow-up for many public health studies and surveillance systems in the US[15] and the availability of SSN for linkage has important implications for analyses using the linked NHIS-NDI data. In particular, differences have been found between participants who provide a SSN vs those who do not[16] and differences in SSN availability by race and ethnicity could differentially affect the ability to link and potentially bias mortality disparities reported with the linked data.[2] Since the evaluation by Lariscy,[9] several changes in the collection of SSNs have greatly affected the number of participants with a SSN available for linkage.[16] For example, beginning in 2002, SSNs were no longer collected from every member of the sampled household and as a result of these changes, the majority of NHIS participants are being linked to the NDI without a SSN. It is unclear how many links are missed (ie, deaths) for those without a SSN and if missing links differ by race and ethnicity.

The objectives of this analysis are to: 1) compare the availability of a SSN for linkage by race and ethnicity; 2) estimate the percentage of expected links missed due to participants not providing SSNs; and 3) assess if differences in estimated missed links creates biases in mortality ratios reported in the NHIS linked mortality data. Objectives 2 and 3 assume that decedents with a SSN reported at interview would be identified in the NDI; although, in practice, a small number are not identified due to reporting errors in one or both data sources.

## Methods

We used 1989-2009 NHIS data linked with death records from the NDI through December 31, 2011. The NHIS is a large continuous survey conducted by NCHS. The NHIS uses a complex survey design, including oversamples of Black, Hispanic and Asian persons. The NHIS data files include sample weights that account for the oversampling and other factors so that estimates are representative of the US civilian non-institutionalized population. The NHIS includes demographic information and basic health characteristics on everyone in the sampled household. Since 1997, one adult and one child within the household are sampled to complete more detailed questionnaires. Additional information on the NHIS, its sample design and weighting methods can be found elsewhere.[17]

Until 2002, the NHIS requested SSNs from everyone in the household. From 2002 forward, SSNs were only requested from a family respondent and/or sample adult and child. In addition, since 2007, only the last four digits of the SSN have been collected to counter the increasing number of participants refusing to provide a full SSN.[16]

The NDI is a centralized database of death records from deaths occurring in the US since 1979.[15] The NDI contains identifying information from the death certificate that can be used to match records from other data sources for epidemiologic studies. The information collected on the death certificate is typically provided by a range of sources, which can include funeral directors, medical examiners, attending physicians, and next-of-kin.

## Data Linkage

We used data from the most recent NCHS linkage between the NHIS and NDI. Matching variables in the probabilistic linkage included: SSN, sex, first and last name, middle initial, father's surname (for women when available), full date of birth, race, state of birth, state of residence and marital status. Linkage methods were modified to accommodate the change from a 9-digit to 4-digit SSN.[12] An evaluation by NCHS found 99.7% agreement with vital status comparing linkage methods with 9- and 4-digit SSNs.[18] In order to be eligible for linkage to the NDI, survey participants could not refuse to have their data linked and needed to provide a minimum level of PII. Each matching variable received a score, which was weighted by its uniqueness in the linkage process, and total scores above a certain threshold were considered a true match and the participant was considered deceased. Potential matches with a score that fell below the threshold were considered false matches and the participant was assumed to be alive. A more detailed description of the linkage process, including the scoring process, is available online.[19]

In the linkage process, a small percentage of potential matches below the score thresholds were manually reviewed and changed to a true match if there was sufficient evidence based on visual inspection of the PII on both files. Since we examined the ability to link by SSN and researchers outside of NCHS who plan to link with the NDI do not have the ability to view identifiable information to conduct manual review, survey participants with deaths determined by manual review were excluded from this analysis (< 2% of all linked deaths). In addition, survey participants with deaths only identified through sources other than the NDI were excluded.

## Study Population

The analysis was conducted among survey participants aged ≥65 years at the time of survey. Results are presented separately for Hispanic, White non-Hispanic, Black non-Hispanic and Asian/Pacific Islander (PI) participants and for these groups combined. Although NHIS currently reports results separately for Asian participants and Native Hawaiian or Other Pacific Islander participants, the combined Asian/PI category is used to be consistent for the entire survey period.[20] There were too few participants who identified as American Indian/Alaskan Native, who selected multiple races, or who were coded as 'Other' to examine separately; participants in these groups were not included in either the separate or combined analyses.

Of the 240,807 NHIS participants aged ≥65 years at interview, 13,088 (5.4%) were ineligible for the linkage either due to not consenting to have their data linked or due to insufficient PII; 3,076 (1.4%) were excluded because their potential linkage underwent manual review or had a source of death other than the NDI; and 3,903 (1.7%) were excluded due to missing data for variables included in the analysis or because the participants were not classified into one of the four race/ethnic groups. There were 220,740 participants available for analysis (White non-Hispanic: 174,428; Black non-Hispanic: 24,529; Hispanic: 17,193; Asian/PI: 4,590).

## Statistical Analysis

All analyses with the linked NHIS-NDI data were conducted with SAS (9.3) and SAS-callable SUDAAN (11.0) to account for the complex survey design and produce weighted estimates. We used chi-square tests to test overall statistical differences of demographic and health characteristics among the race and ethnic groups. We also compared the percent of records linked by race and ethnicity and by the availability of a SSN.

To estimate the percent of missed linked records due to participants not providing a SSN (either due to not having it, not knowing it, refusing to provide it, or not being asked), we used preventive fraction methods.[21] The prevented fraction is defined as the proportion of a hypothetical total for a particular outcome that was prevented due to exposure to a protective factor.[22] In this case, we considered not having a SSN as protective against linking with the NDI and therefore, links were "prevented" due to "exposure" to not having a SSN. This method assumes that some survey participants without SSNs died and did not link with the NDI but would have linked if they had provided a SSN. The preventive fraction among the exposed is calculated as 1 minus the relative risk.

In this analysis, the relative risk of linking with the NDI was approximated based on Cox proportional hazards models from PROC SUR-

VIVAL in SUDAAN accounting for potential differences in follow-up times and adjusting for factors related to the availability of a SSN and to mortality. First, the proportional hazard assumption was examined using Kaplan-Meier curves. Then, using participants with a reported SSN as the unexposed group (the reference group), we calculated the hazard ratio (HR) of linking with the NDI for those missing a SSN. Models were adjusted for: survey year (1989-1996; 1997-2001; 2002-2009); sex; categorized age at interview (65-74; 75-84; ≥85); respondent-rated health (fair/poor vs good/very good/excellent); education level (<high school [HS] degree; HS graduate/post-HS education; Bachelor's degree or higher); and US nativity (US born; born outside the US) and were run for the four race/ethnic groups combined and separately. Race/ethnicity was included as a control variable in the combined model. Follow-up time was based on time from interview to death or end of the follow-up period (December 31, 2011).

Using the HR, the estimated percent of missed links among those without a SSN (exposed) was calculated as 1-HR. To estimate the percent of missed links overall and among each race group, we multiplied the percent of links missed among those without a SSN by the percent of participants with no SSN (p) (ie, p*(1-HR)).[21] In order to calculate confidence intervals (CI) for p*(1-HR), we used bootstrap sampling. We produced 1,000 samples using the rescaling bootstrap method for complex surveys, as described by Rao et al (1992) and implemented using SAS code adapted from the work of Cheng, et al (2008), to obtain estimates and their respective percentile-based 95% confidence intervals.[23,24] This method involves rescaling the original survey weights in order to produce bootstrap weights that are then used to estimate some parameter of interest. In our application, the parameters of interest were the preventive fractions among the exposed and the overall preventive fractions. Preventive fractions were considered significantly different between race and ethnic groups if bootstrap 95% CI calculated for the differences in these estimates between groups did not include zero.

To further examine if the estimated percent of missed links by race/ethnicity biases mortality ratios, we used the estimated percent of missed links to calculate a corrected percent linked (ie, the percent we would have expected to link if all of the survey participants had had a SSN). The corrected percent linked was calculated as the following:

Corrected Percent Linked = Observed Percent Linked / (1-Overall Percent Missed)

Finally, we calculated mortality ratios by race/ethnicity obtained using the observed and corrected percent linked. As an approximate comparison, we included race/ethnicity-specific age-adjusted mortality rate ratios for adults aged ≥65 years calculated from the NCHS Compressed Mortality File for 1999-2011.[25] The average rates over this time-period were age-adjusted to the 2000 US standard population and ratios were calculated with the White non-Hispanic population as the reference group.

## Results

The number of participants in the study population along with the weighted distribution of characteristics by race/ethnicity are presented in Table 1. There were significant differences in characteristics by race/ethnicity (all chi-square tests were statistically significant at the P<.05 level). There was a higher percentage of women compared with men in each race/ethnic group, with the highest percentage among Black non-Hispanic participants. White non-Hispanic participants were older than the other groups with significantly more participants aged ≥75 years. There were substantial differences between the percentages of participants born outside of the US by race/ethnicity. While less than 6% of White and Black non-Hispanic participants across cohorts were born outside the US, close to 60% and 70% of Hispanic and Asian/PI participants, respectively, were born outside the US. The highest percentage of participants born outside the US was among Asian/PI participants (71.8%). Respondent-rated health tended to be comparable between Hispanic and Black non-Hispanic participants and between Asian/PI and White non-Hispanic participants. A higher percentage of Hispanic and Black non-Hispanic participants reported fair/poor health compared with Asian/PI and White non-Hispanic participants. Similarly, education levels were more comparable between the same groups. However, Hispanic participants had the highest percentage with less than a high school education (60.5%).

Survey years were divided to co-

**Table 1. Weighted distribution of selected characteristics by race and ethnicity among 1989-2009 National Health Interview Survey participants aged ≥65 years at interview included in the linkage with the National Death Index (follow-up through 2011)**

| | Hispanic weighted % n=17,193 | Asian/Pacific Islander weighted % n=4,590 | Black non-Hispanic weighted % n=24,529 | White non-Hispanic weighted % n=174,428 |
|---|---|---|---|---|
| Overall, row percent | 5.2 | 2.0 | 8.3 | 84.5 |
| Observed linked, deceased[a] | 35.6 | 27.7 | 49.2 | 50.1 |
| Female [a] | 57.8 | 57.0 | 60.5 | 57.3 |
| Age group [a] | | | | |
| 65-74 years | 63.5 | 62.6 | 60.4 | 54.8 |
| 75-84 years | 29.5 | 29.9 | 31.0 | 35.1 |
| 85 and older | 7.0 | 7.5 | 8.6 | 10.1 |
| Born outside US [a], vs native born | 58.5 | 71.8 | 4.8 | 5.5 |
| Fair/poor health [a], vs excellent, very good, or good health | 37.0 | 25.2 | 41.0 | 24.7 |
| Education level [a] | | | | |
| < High school (HS) | 60.5 | 29.4 | 52.2 | 27.4 |
| HS graduate/ GED / some college | 31.3 | 43.1 | 39.1 | 55.9 |
| ≥ Bachelor's degree | 8.2 | 27.5 | 8.7 | 16.7 |
| Survey year [a] | | | | |
| 1989-1996 | 27.3 | 24.8 | 36.6 | 37.6 |
| 1997-2001 | 23.7 | 21.6 | 22.3 | 22.9 |
| 2002-2009 | 49.1 | 53.7 | 41.1 | 39.6 |

a. Chi-square tests by race/ethnicity statistically significant at the α=.05 level.

incide with changes in sample design and data collection. Consistent with changes in oversampling of the Hispanic and Asian/PI populations, a higher percentage of Hispanic and Asian/PI participants were included from the later survey time-periods compared with White and Black non-Hispanic participants (Table 1). For example, 49.1% of Hispanic and 53.7% of Asian/PI participants were interviewed in 2002-2009, compared with 39.6 of White non-Hispanic and 41.1% of Black non-Hispanic participants.

Table 2 presents the weighted distribution of SSN availability by race/ethnicity along with the corresponding observed percent linked, adjusted HR of linking without a SSN, estimated percent of links missed among the exposed (those missing a SSN), and overall estimated percent missed. Hispanic and Asian/PI participants were less likely to have a SSN compared with White and Black non-Hispanic participants (P<.01 for both). Asian/PI participants had the highest percentage without a SSN, with close to 60% (57.7%). In each race/ethnic group, the observed percent linked significantly differed by the availability of a SSN. After adjustment, the HR for linking without a SSN ranged from .66 among Hispanic participants to .77 among White non-Hispanic participants. With the preventive fraction methods (ie, 1-HR), this translated to an estimated 23%-34% of links missed in the different race/ethnic groups among participants who did not provide a SSN. Compared with White non-Hispanic participants, the difference in the percent missed links among those without a SSN was significantly higher for Hispanic (10.9 95%CI 6.6, 15.5) and Black non-Hispanic participants (6.3 95% CI 2.7, 9.6) but not for Asian/PI participants (8.6 95% CI -1.7, 17.6) (differences not presented in Table 2).

When we multiplied the percent of participants missing a SSN within each race/ethnic group by the estimated percent of links missed among the exposed, the overall estimated percent of missed links was highest among Asian/PI (18.5%) and Hispanic (18.1%) participants and lowest among White non-Hispanic (9.9%) participants (Table 2). However, the CI around the estimate for Asian/PI participants was wide. Differences in the percent of missed links were significantly higher for Asian/PI (8.6 95%CI 2.8, 14.0), Hispanic (8.2 95%CI 5.9, 10.5) and Black non-Hispanic participants (3.5 95%CI 1.8, 5.0) compared with White non-Hispanic participants. For these four groups combined, it was estimated that 10.7% of links

**Table 2. Weighted distribution of 1989-2009 National Health Interview Survey participants aged ≥65 years at interview by availability of a Social Security Number (SSN) and corresponding % linked, adjusted hazard ratio (HR) and estimated % of links missed due to missing SSN**

| | % distribution | % linked | Adjusted[a] HR for linking with the NDI (HR) | 95% CI | % of links missed among those without a SSN (1-HR) | 95% CI | % of links missed within race/ethnic group overall P[a](1-HR) | 95% CI |
|---|---|---|---|---|---|---|---|---|
| All groups combined | | | | | | | | |
| With SSN | 56.8 | 59.8 | 1.0 (Ref) | | (unexposed) | | | |
| Missing SSN | 43.2 | 34.3 | .75 | .74, .76 | 24.8 | 23.6, 25.9 | 10.7 | 10.2, 11.2 |
| Hispanic | | | | | | | | |
| With SSN | 47.4 | 48.9 | 1.0 (Ref) | | (unexposed) | | | |
| Missing SSN | 52.6 | 23.5 | .66 | .61, .70 | 34.4 | 30.4, 38.7 | 18.1 | 15.9, 20.4 |
| Asian/Pacific Islander | | | | | | | | |
| With SSN | 42.3 | 39.6 | 1.0 (Ref) | | (unexposed) | | | |
| Missing SSN | 57.7 | 18.9 | .68 | .59, .78 | 32.0 | 21.9, 41.0 | 18.5 | 12.7, 23.9 |
| Black non-Hispanic | | | | | | | | |
| With SSN | 55.1 | 60.3 | 1.0 (Ref) | | (unexposed) | | | |
| Missing SSN | 44.9 | 35.7 | .70 | .67, .73 | 29.8 | 26.5, 33.0 | 13.4 | 11.9, 14.8 |
| White non-Hispanic | | | | | | | | |
| With SSN | 57.9 | 60.7 | 1.0 (Ref) | | (unexposed) | | | |
| Missing SSN | 42.1 | 35.5 | .77 | .75, .78 | 23.5 | 22.1, 24.8 | 9.9 | 9.3, 10.5 |

a. Adjusted for age group, sex, education level, respondent-rated health status, US nativity, and survey years. The model among all groups also adjusted for race/ethnicity.

were missed due to lack of SSNs.

In the next step, we corrected for links we presumed would have been captured if a SSN was available (Table 3). Despite finding higher estimates of missed links among Hispanic and Asian/PI participants, the absolute percentage point increase from the observed percent to the corrected percent linked was modest. The corrected percent linked ranged from 4.6 points higher among White non-Hispanic participants to 6.8 points higher among Black non-Hispanic participants. As a result, there was a modest change in the ratios for each race/ethnicity relative to White non-Hispanic participants. Compared with the national mortality rate ratios averaged over 1999-2011 with the White non-Hispanic group as the reference group, the corrected ratios were very similar for Hispanic and Asian/PI participants. However, the corrected ratio for Black non-Hispanic participants (ratio=1.02) remained lower than the national rate ratio (ratio=1.13).

## Discussion

In this analysis of NHIS-NDI linked data, we found evidence that a nontrivial percent of links (deaths) may be missed among survey participants without a SSN available for linkage. A disproportionate number of links appeared to be missed for Hispanic and Asian/PI participants compared with the other groups. While the difference was not as large, the estimated percent of missed links was higher among Black non-Hispanic participants compared with White non-Hispanic participants, as well. Despite finding differences in the ability to link compared with White non-Hispanic participants, correcting for the missed links did not have a large impact on the ratio of the adjusted percent linked between the non-Hispanic White compared with the others groups. As a result, missed links due to lack of a SSN likely has only a modest impact on mortality disparities in this data.

Similar to the published literature, we found that Hispanic participants were less likely to have a SSN available for linkage compared with White and Black non-Hispanic participants.[10,11,13] However, because not having a SSN in this analysis is due to the combination of participants not knowing, not having, not being asked (per study protocol) or unwilling to provide it, the percentages in this analysis are substantially higher. Consistent with earlier studies,[9,11,13,26] the higher percentage of estimated missed links provides further evidence

**Table 3. Actual and corrected % linked and corresponding ratios compared to White Non-Hispanic participants among 1989-2009 National Health Interview Survey Participants aged >65 years at interview linked with the National Death Index through 2011. National mortality rate ratios are included for comparison**

| | Observed % linked | Observed ratio | Corrected % linked | Corrected ratio | National mortality rate ratios[a] 1999-2011 |
|---|---|---|---|---|---|
| Hispanic | 35.6 | .71 | 42.1 | .77 | .75 |
| Asian/Pacific Islander | 27.7 | .55 | 32.8 | .60 | .60 |
| Black non-Hispanic | 49.2 | .98 | 56.0 | 1.02 | 1.13 |
| White non-Hispanic | 50.1 | 1.0 (Ref) | 54.7 | 1.0 (Ref) | 1.0 (Ref.) |

a. Age-adjusted mortality rate ratios for adults aged ≥65 years calculated from the NCHS Compressed Mortality File for 1999-2011. Rates were age-adjusted to the 2000 US standard population.

that the linkage process may be less effective for the Hispanic population. However, this analysis contributes to the literature by quantifying the expected missed links due to missing SSN, finding that the estimated missed links were almost 90% higher among Hispanic and Asian groups compared with White non-Hispanic, groups. Importantly, this method could be used by other researchers using linkage with the NDI or other death certificate data to estimate differential linkage in their study.

In addition to SSN, there may be other factors that contribute to differential linkage that this analysis did not capture. As mentioned previously, Hispanic and Asian/PI participants can be more difficult to link because of naming conventions and there is some evidence that those born outside the US are less likely to know their exact date of birth due to lack of birth registration.[27] Furthermore, there is evidence that some immigrants may move back to their country of origin where their death may not be recorded by the NDI – a phenomenon referred to as "salmon bias."[28] Although we controlled for US nativity, we did not specifically examine the ability to link by US nativity in this analysis. Another limitation of this analysis is our inability to examine subgroups of the Hispanic and Asian/PI participants. Although we found evidence of heterogeneity among Hispanic and Asian/PI subgroups in preliminary analyses, we did not have sufficient numbers to examine them separately for these analyses.

This analysis had a number of important strengths. The large study population allowed us to include results for Asian/PI participants, which many studies have not been able to assess. In addition, because of the content of the NHIS, we were able to control for US nativity, education level and respondent-reported health in our survival models, all of which are significantly associated with mortality and differ by race and ethnicity.[29,30] And, by producing separate estimates by race/ethnicity and controlling for some baseline factors, we were able to separate differences in the ability to link from other differences in mortality between groups. However, residual confounding is possible, which would, in turn, bias the preventive fraction estimates. In addition, estimates are likely affected by false links that may differ across groups or missing links due to errors in reported SSN, which are not addressed in this study.

As the NDI is commonly used for passive mortality follow-up, this analysis demonstrates the importance of having a SSN available for linkage. Although SSNs were missing for a variety of reasons in this study population and the percentage of participants missing a SSN was higher for all the race and ethnic groups compared with other studies, the differences in the ability to link across groups observed in this analysis still demonstrate the potential for bias in studies linking with death certificate data. From a broader perspective, with increasing interest in Big Data and data linkage, there is added importance for researchers to consider the potential impact that differences in the ability to link records across groups might have on inferences from their analyses. With data coming from multiple sources, there will be differences in availability, quality, and format of unique identifiers, which could disproportionately affect minority populations. Despite some reassurance that the differential linkage from SSN availability in this analysis did not have a large impact on our ratios, researchers conducting analyses of racial/ethnic mortality disparities using the NDI or other health disparities using linked records (eg, medical records), need to be cognizant of the potential for differences in the ability to link to contribute to their results. Furthermore, preventive fraction methods might be a tool that researchers can use to assess, and possibly correct for, potential bias in their own studies.

## Acknowledgments

The authors would like to thank Dr. Makram Talih for his valuable feedback and suggestions on this manuscript and Dr. Milton Kotelchuck for his insightful thoughts on how to estimate missed links.

The conclusions, findings, and opinions expressed by authors contributing to this journal do not necessarily reflect the official position of the U.S. Department of Health and Human Services, the Public Health Service, the Centers for Disease Control and Prevention, or the authors' affiliated institutions.

## Conflict of Interest

No conflicts of interest to report.

## Author Contributions

Research concept and design: Miller, McCarty, Parker; Data analysis and interpretation: Miller, McCarty; Manuscript draft: Miller, Parker; Statistical expertise: Miller, McCarty; Supervision: Parker

## References

1. Weir HK, Johnson CJ, Mariotto AB, et al. Evaluation of North American Association of Central Cancer Registries' (NAACCR) data for use in population-based cancer survival studies. *J Natl Cancer Inst Monogr*. 2014;2014(49):198-209. http://dx.doi.org/10.1093/jncimonographs/lgu018. PMID:25417233.
2. Centers for Disease Control and Prevention. Use of the National Death Index in Health Research. http://www.cdc.gov/nchs/data/ndi/ndi_bibliography_health_research.pdf. Accessed 4/28/2016.
3. Capistrant BD, Gilsanz P, Moon JR, Kosheleva A, Patton KK, Glymour MM. Does the association between depressive symptoms and cardiovascular mortality risk vary by race? Evidence from the Health and Retirement Study. *Ethn Dis*. 2013;23(2):155-160. PMID:23530295.
4. Borrell LN, Dallo FJ, Nguyen N. Racial/ethnic disparities in all-cause mortality in U.S. adults: the effect of allostatic load. *Public Health Rep*. 2010;125(6):810-816. PMID:21121226.
5. Borrell LN, Lancet EA. Race/ethnicity and all-cause mortality in US adults: revisiting the Hispanic paradox. *Am J Public Health*. 2012;102(5):836-843. http://dx.doi.org/10.2105/AJPH.2011.300345. PMID:22493998.
6. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. *Ann Epidemiol*. 2002;12(7):462-468. http://dx.doi.org/10.1016/S1047-2797(01)00285-X. PMID:12377423.
7. Markides KS, Coreil J. The health of Hispanics in the southwestern United States: an epidemiologic paradox. *Public Health Rep*. 1986;101(3):253-265. PMID:3086917.
8. Xu J, Murphy SL, Kochanek KD, Bastian BA. *Deaths: Final Data for 2013. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics*. National Vital Statistics System. 2016;64(2):1-119.
9. Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *J Aging Health*. 2011;23(8):1263-1284. http://dx.doi.org/10.1177/0898264311421369. PMID:21934120.
10. Patel KV, Eschbach K, Ray LA, Markides KS. Evaluation of mortality data for older Mexican Americans: implications for the Hispanic paradox. *Am J Epidemiol*. 2004;159(7):707-715. http://dx.doi.org/10.1093/aje/kwh089. PMID:15033649.
11. Pinheiro PS, Williams M, Miller EA, Easterday S, Moonie S, Trapido EJ. Cancer survival among Latinos and the Hispanic Paradox. *Cancer Causes Control*. 2011;22(4):553-561. http://dx.doi.org/10.1007/s10552-011-9727-6. PMID:21279543.
12. Palloni A, Arias E. Paradox lost: explaining the Hispanic adult mortality advantage. *Demography*. 2004;41(3):385-415. http://dx.doi.org/10.1353/dem.2004.0024. PMID:15461007.
13. Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records--accuracy and sources of bias. *J Clin Epidemiol*. 2004;57(1):21-29. http://dx.doi.org/10.1016/S0895-4356(03)00250-6. PMID:15019007.
14. Hoeffel E, Rastogi S, Kim M. The Asian Population: 2010 - U.S. Census Briefs. 2012.
15. Centers for Disease Control and Prevention. National Death Index. http://www.cdc.gov/nchs/ndi.htm. Accessed 4/28/2016.
16. Miller DM, Gindi RM, Parker JP. Trends in record-linkage refusal rates: characteristics of National Health Interview Survey participants who refuse record-linkage. Paper presented at: American Statistical Association Joint Statistical Meetings. 2011.
17. Center for Disease Control and Prevention. National Health Interview Survey. 2015; http://www.cdc.gov/nchs/nhis.htm. Accessed 2/27/2015.
18. McCarty F. Evaluation of the 4-Digit Social Security Number Algorithm Used for the 2011 Linked Mortality Files. Paper presented at: Centers for Disease Control and Prevention National Conference on Health Statistics, 2015; Bethesda, MD.
19. Centers for Disease Control and Prevention. NCHS 2011 Linked Mortality Files Matching Methodology, September, 2013. http://www.cdc.gov/nchs/data/datalinkage/2011_linked_mortality_file_matching_methodology.pdf.
20. Centers for Disease Control and Prevention. Race and Hispanic Origin Information. http://www.cdc.gov/nchs/nhis/rhoi.htm. Accessed 2/27/2015.
21. Kleinbaum DG, Sullivan KM, Barker ND. *ActivEpi Companion Textbook*. New York: Springer Science & Business Media; 2013. http://dx.doi.org/10.1007/978-1-4614-5428-1.
22. Last JM. *A Dictionary of Epidemiology*. 3rd ed. Oxford, UK: Oxford University Press; 1995.
23. Cheng NF, Han PZ, Gansky SA. Methods and software for estimating health disparities: the case of children's oral health. *Am J Epidemiol*. 2008;168(8):906-914. http://dx.doi.org/10.1093/aje/kwn207. PMID:18779387.
24. Rao JK, Wu CJ, Yue K. Some recent work on resampling methods for complex surveys. *Surv Methodol*. 1992;18:209-217.
25. National Center for Health Statistics Compressed Mortality File (CMF) on CDC WONDER Online Database. 2014. Accessed 4/29/2015.
26. Leiss JK, Giles D, Sullivan KM, Mathews R, Sentelle G, Tomashek KMUS. U.S. Maternally linked birth records may be biased for Hispanics and other population groups. *Ann Epidemiol*. 2010;20(1):23-31. http://dx.doi.org/10.1016/j.annepidem.2009.09.003. PMID:20006273.
27. UNICEF Innocenti Research Centre. Birth registration: right from the start. *Innocenti Digest*; 2002:9.
28. Pablos-Méndez A. Mortality among Hispanics. *JAMA*. 1994;271(16):1237. http://dx.doi.org/10.1001/jama.1994.03510400023017. PMID:8031363.
29. Lariscy JT, Hummer RA, Hayward MD. Hispanic older adult mortality in the United States: new estimates and an assessment of factors shaping the Hispanic paradox. *Demography*. 2015;52(1):1-14. http://dx.doi.org/10.1007/s13524-014-0357-y. PMID:25550142.
30. Ross CE, Masters RK, Hummer RA. Education and the gender gaps in health and mortality. *Demography*. 2012;49(4):1157-1183. http://dx.doi.org/10.1007/s13524-012-0130-z. PMID:22886759.