The Impact of Follow-up Type and Missed Deaths on Population-Based Cancer Survival Studies for Hispanics and Asians

Paulo S. Pinheiro, Cyllene R. Morris, Lihua Liu, Timothy J. Bungum, Sean F. Altekruse

Correspondence to: Paulo S. Pinheiro, PhD, Epidemiology and Biostatistics, School of Community Health Sciences, University of Nevada Las Vegas, 4505 S Maryland Parkway, Las Vegas, NV 89154 (e-mail: paulo.pinheiro@unlv.edu).

Background
The accuracy of cancer survival statistics relies on the quality of death linkages and follow-up information collected by population-based cancer registries. Methodological issues on survival data by race-ethnicity in the United States, in particular for Hispanics and Asians, have not been well studied and may undermine our understanding of survival disparities.

Methods
Based on Surveillance, Epidemiology, and End Results (SEER)-18 data, we analyzed existing biases in survival statistics when comparing the four largest racial-ethnic groups in the United States, whites, blacks, Hispanics and Asians. We compared the “reported alive” method for calculation of survival, which is appropriate when date of last alive contact is available for all cases, with the “presumed alive” method used when dates of last contact are unavailable. Cox regression was applied to calculate the likelihood of incomplete follow-up (those with less than 5 years of vital status information) according to racial-ethnic group and stage of diagnosis. Finally, potentially missed deaths were estimated based on the numbers of cases with incomplete follow-up for highly fatal cancers.

Results
The presumed alive method overestimated survival compared with the reported alive method by as much as 0.9–6.2 percentage points depending on the cancer site among Hispanics and by 0.4–2.7 percentage points among Asians. In SEER data, Hispanics and Asians are more likely to have incomplete follow-up than whites or blacks. The assumption of random censoring across race-ethnicity is not met, as among non-white cases, those who have a worse prognosis are more likely to have incomplete follow-up than those with a better prognosis (P < .05). Moreover, death ascertainment is not equal across racial-ethnic groups. Overall, 3% of cancer deaths were missed among Hispanics and Asians compared with less than 0.5% among blacks and whites.

Conclusions
Cancer survival studies involving Hispanics and Asians should be interpreted with caution because the current available data overtly inflates survival in these populations. Censoring is clearly nonrandom across race-ethnicity meaning that findings of Hispanic and Asian survival advantages may be biased. Problematic death linkages among Hispanics and Asians contribute to missing deaths and overestimated survival. More complete follow-up with at least 5 years of information on vital status as well as improved death linkages will decisively increase the validity of survival estimates for these growing populations.


The accurate estimation of cancer survival across racial and ethnic groups is an important aspect of cancer surveillance. It relies on completeness of follow-up to identify cancer deaths. In cancer surveillance, there are two different methods of follow-up. Active follow-up is defined as any type of follow-up that involves contact with the patient, his/her next-of-kin or his/her physician to ascertain vital status. Passive follow-up involves obtaining follow-up data through linkages without contact with the patient or his/her physician.

Cancer registration in the United States is administered by one of two programs that use different types of follow-up. The first, the National Cancer Institute Surveillance, Epidemiology and End Results (SEER), covers 28% of the US population and employs resource intensive activities to capture the last date of contact for each cancer case. Cases are matched to national and statewide databases such as the National Death Index, Social Security, Medicare and Medicaid, hospital discharge, as well as records of contact with physicians, pathology labs, and hospital registries. SEER registries also perform data linkages with voter registration and driver license data, and some use information obtained from research studies to supplement registry data. SEER standards require that their registries report vital status and date of last contact that is current within 22 months of the date of their annual data submission for a minimum of 95% of all registered cancer patients, living and deceased (1). The follow-up information can be from either linkages or direct contact with patients, so SEER registries perform a combination of active and passive follow-up. In spite of this, in this paper we refer to the follow-up carried out by SEER as “active”
follow-up because a date of last contact is available for all cases. Because of the availability of this last date of contact, the traditional method of survival analysis with SEER data is the “reported or documented alive” method, which relies on the reported date of death and the date of alive last contact if there is no date of death available. This method censors patients at the date of last alive contact. Researchers who perform survival analyses often use SEER data because of its high quality. Moreover, this was the only source of population level survival data available until recently.

The second US cancer surveillance program is the National Program for Cancer Registries (NPCR). The NPCR was established by the Centers for Disease Control and Prevention and covers the balance of the US population. NPCR registries conduct passive follow-up by matching cancer cases with in-state death lists and the National Death Index. However, NPCR registries do not systematically collect a date of last contact for living cases. As a result, NPCR survival statistics are calculated using what is known as the “presumed alive” method. This method assumes that if a person is not found to be dead, then he/she is alive on the most recent date covered by the National Death Index. Presumed alive survival statistics appear less commonly in the literature (2–5), partly because NPCR survival data have only recently become available. Both Centers for Disease Control and Prevention and the North American Association of Central Cancer Registries now recognize passive follow-up as a suitable approach for some types of population-based survival calculations (6).

In the United States, Hispanics, and Asians and Pacific Islanders (APIs) are the fastest growing racial and ethnic groups, comprising 22% of the US population (7). The effects of active versus passive follow-up and their corresponding methods of estimating cancer survival have not been thoroughly examined among these groups. We hypothesized that active and passive follow-up procedures are not equally effective across the four major US racial-ethnic groups: whites, blacks, Hispanics, and APIs.

In this study, five-year survival proportions were obtained and compared for these groups using the two described methods, reported alive and the presumed alive, that is, by considering the date of last contact (active follow-up conducted by SEER) and by assuming the patient was alive if not reported dead within 5 years of diagnosis (passive follow-up conducted in NPCR registries), respectively. In addition, the completeness of SEER “active” follow-up was compared across racial-ethnic groups (whites, blacks, Hispanics, and APIs) by stage at diagnosis and by racial-ethnic subgroups (eg, Mexican, Chinese, and so on) to assess potential biases across racial-ethnic populations.

Methods
Study Data
De-identified data from 18 SEER registries were used for the analyses (November 2011 submission) (8). Active follow-up extended to December 31, 2009. Non-Hispanic white, non-Hispanic black, non-Hispanic API, and Hispanic subjects were categorized into mutually exclusive groups as whites, blacks, APIs, and Hispanics, respectively. Hispanics were identified via the North American Association of Central Cancer Registries Hispanic Identification Algorithm v2.2.1, regardless of race (9). Non-Hispanic cases were classified according to primary race. American Indian/Alaska Natives were not included in the analyses because of small numbers.

The SEER site classification and the SEER historic stage variables were used to categorize cancer site and stage (10,11). Because prostate cancer stage combined localized and regional cases, the final covariate was defined in three categories—local/regional, distant, and unknown stage. Cases identified solely by death certificates were excluded from the analyses. This study was approved by the University of Nevada Las Vegas Institutional Review Board.

Statistical Analyses
Several analyses were carried out to complete the study objectives. In the first analysis, cases diagnosed from 2000–2008 were used to calculate and compare five-year observed survival using both the reported alive and the presumed alive methods by racial-ethnic group. To calculate survival time based on the reported alive method, the SEER dates of last contact for cases with no reported death were used. To simulate a scenario of 100% passive follow-up for the presumed alive method, the date of last contact for non-dead cases was set to December 31, 2009. As a practical example illustrating the difference between the reported alive and the presumed alive methods, if a non-dead patient was diagnosed in 2003 and had a last date of alive contact at 4 years after diagnosis, he or she would be censored as alive at 4 years with the reported alive method, but would have contributed more than 5 years with the presumed alive method. Observed survival for the first primary cancer was age-adjusted using the international cancer survival standards (12). Survival was calculated for nine cancer sites including four leading sites (prostate, female breast, lung and bronchus, and colorectal) and five sites with less favorable prognoses. Cancer sites with less favorable prognoses were the esophagus, gallbladder, liver and intrahepatic bile duct, pancreas, and stomach. These diagnoses as well as lung and bronchus cancer all have an age-adjusted five-year observed survival rate of less than 25% (13).

In the second analysis, we wanted to assess the completeness of vital status information in SEER registries, or in other words, whether SEER “active” follow-up was able to gather dates of last contact equally across the four racial-ethnic groups. We set a cutoff of 5 years (60 months) after diagnosis because 5 years is the conventional period for survival statistics. All the deceased regardless of their date of death and those alive with more than 5 years of follow-up were considered to have complete follow-up. If a non-dead patient did not have at least 5 years of information as alive, then he/she was considered to have incomplete follow-up or to be “lost to follow-up.” We included cases of leading cancer sites (prostate, female breast, lung and bronchus, and colon and rectum) diagnosed between 2000 and 2003. The years of diagnosis, 2000–2003, allowed for the incorporation of SEER cases to potentially contribute 5 years of follow-up. The distributions of these “lost to follow-up” cases were examined by racial-ethnic group. To study the influence of race-ethnicity and stage of diagnosis on incomplete follow-up, we used Cox regression to model the time to being lost to follow-up (14,15). In these analyses, being lost to follow-up was the event of interest and deaths before 60 months (5 years) were censored at the time of death. Deaths occurring after 60 months of follow-up and cases with a date of last alive contact more than 60 months after diagnosis were censored at 60 months. The Cox regression models were adjusted for age, stage
at diagnosis, diagnosis year, SEER registry, cancer site, and gender and were run first for all racial-ethnic groups combined, and then separately by racial-ethnic group. Hazard ratios values greater than 1.0 indicate a greater likelihood of being lost to follow-up (having incomplete follow-up) in comparison to the reference group.

In our final analysis, we wanted to ascertain whether SEER “active” follow-up was able to gather information on deaths equally across the four racial-ethnic groups and, within Hispanics and APIs, across the different subgroups they comprise. For this, the percent of deaths missed in the death linkage process was estimated by race-ethnic group based on the proportions of those lost to follow-up (according to the 60-months cutoff) of highly fatal diagnoses. We defined highly fatal diagnoses as distant stage cancers of four sites for which survival after 5 years was less than 3%: liver and intrahepatic bile duct, gallbladder, pancreatic, and lung and bronchus (13). This subset of cases provided enough observations in every racial-ethnic group for meaningful analysis. Cases with highly fatal cancer diagnoses between 2000 and 2003 classified as alive with last contact before 60 months were classified as “missed deaths,” on the assumption that cases diagnosed with these cancers and lost to active follow-up before 60 months were likely to be missed deaths regardless of race-ethnicity. We estimated the proportion of missed deaths based on the quotient of missed deaths over the sum of missed and recorded deaths before 60-months postdiagnosis. In addition to the major racial-ethnic groups (whites, blacks, Hispanics, and APIs), Mexican, Puerto Rican, Cuban, South and Central American, and other Hispanic subgroups were examined as were Chinese, Japanese, Hawaiian, Filipino, Korean, Vietnamese, South Asians, and other APIs. The South Asian subgroup included Indian, Pakistani, and Asian Indian categories.

### Results

Between 2000 and 2008, there were 1,471,789 cases diagnosed with leading cancer sites (female breast, colorectal, lung and bronchus, and prostate) and 181,515 cases diagnosed with less favorable prognosis cancers (esophagus, stomach, liver and intrahepatic bile duct, gallbladder, and pancreas). Survival proportions for all cancer sites and across all racial/ethnic groups were consistently higher using the presumed alive method rather than the reported alive method (Table 1). Survival changed less for whites and for blacks than for other groups. Among Hispanics, the net differences in percent survival between the two methods varied from 0.7 percentage points for prostate cancer to 6.2 for gallbladder cancer ($P < .05$). Among APIs, the differences ranged from 0.4 percentage points for prostate cancer to 2.7 for gallbladder cancer. For lung and bronchus, liver and intrahepatic bile duct, and pancreatic cancers among both Hispanics and APIs, and for gallbladder and stomach cancers among Hispanics only, differences in survival were statistically significant between the two methods.

For leading cancers, the percent of last alive contact before 60 months (“lost to follow-up”) within 5 years of diagnosis (Table 2) was higher among APIs (3.4%) and Hispanic (3.7%) than among white (0.7%) and black (1.1%) cases. The number of cases “lost to follow-up” per 10,000 person-months was higher among Hispanics (8.4) and APIs (7.7) than blacks (2.7) and whites (1.7). Median time to lost follow-up was shorter among Hispanics and APIs (21 and 25 months) than among whites and blacks (33 and 29 months, respectively). Higher rates of lost to follow-up were seen among Hispanics and APIs compared with blacks and whites with less favorable diagnoses and highly fatal diagnoses. In the less favorable diagnosis cohort, median time to lost follow-up ranged from 4 months among APIs to 11 months among whites. For the highly fatal diagnosis cohort, median time to lost follow-up ranged from 3.5 months among blacks to 6 months among whites.

Predictors of “lost to follow-up” were studied among 777,641 cases diagnosed between 2000 and 2003 with leading cancers: female breast, colorectal, lung and bronchus, and prostate (Table 3). Blacks, Hispanics, and APIs had significantly higher likelihood of “lost to follow-up” than the reference category, whites (Table 3). After adjustment for covariates, blacks were 43% more likely than whites to have less than 5 years of follow-up and Hispanics and APIs were 3.7 and 4.1 times more likely to be lost to follow-up, respectively. Overall, those with distant stage diagnoses were 39% more likely to be lost to follow-up than those diagnosed at localized/regional stage (Table 3). The impact of stage of diagnosis differed by racial-ethnic group. For whites, diagnosis at distant stage did not affect risk of lost follow-up compared with diagnosis at localized/regional stage. Among minorities this was not the case. Hispanics diagnosed at distant stage were 69% more likely to be lost to follow-up compared with Hispanics with localized/regional stage cancers. APIs with distant stage cancers were 77% more likely to be lost to follow-up compared with APIs with localized/regional cancers. Blacks diagnosed with distant stage cancer were 36% more likely to be lost to follow-up compared with blacks with localized/regional stage cancers.

Finally, analyses of 115,388 cases with highly fatal cancer diagnoses are presented in Table 4. This cohort was comprised of distant stage liver and intrahepatic bile duct, distant stage gallbladder, distant stage pancreas, and distant stage lung and bronchus cancer cases diagnosed between 2000 and 2003. The proportions of likely missed deaths differed by racial-ethnic group. Whites, blacks, and some Hispanic subgroups (Puerto Ricans), as well as some API subgroups (Japanese and Hawaiian) had low proportions of missed deaths, less than 1%. The percent of likely missed deaths was 2.8% among all Hispanics combined and 2.6% among all APIs combined. Within Hispanic and Asian groups, the percent of likely missed deaths varied substantially by subgroup. High percentages of likely missed deaths were observed among South and Central Americans (9.4%), South Asians (6.1%), Mexicans (4.7%), Filipinos (4.2%), and Chinese (2.8%).

### Discussion

We compared survival estimates across racial-ethnic groups obtained from two possible methods: the reported alive method, which uses date of last contact, and the presumed alive method, which does not. For all cancer sites and racial-ethnic groups, the presumed alive method yielded higher survival estimates. The excess was marginal for most cancer sites and for two large, mostly US-born populations: blacks and whites (16). However, for Hispanics and APIs, the observed differences between the two methods were pronounced and in the less favorable diagnosis cohort (cancers of the lung, pancreas, and so on), survival estimates...
obtained from the two methods differed significantly. Overall, this analysis showed that survival proportions for Hispanics and APIs using the presumed alive method (and respective 100% passive follow-up) should be interpreted with caution, as they may be inflated especially for cancers with less favorable prognoses. If passive follow-up and the presumed alive method result in overestimated survival for Hispanics and APIs, our next question was if the SEER “active” follow-up and the respective reported alive method allow for accurate survival comparisons across racial-ethnic groups. Previous studies have revealed potential biases in survival analyses related to losing cases to follow-up (17–19). When we used SEER data (“active” follow-up), we found that Hispanics and APIs were more likely to be “lost to follow-up” (have <5 years of contact after diagnosis) compared with whites and blacks after adjusting for demographic and case attributes. Moreover, “lost to follow-up” was associated with stage at diagnosis for some groups, but not for others after stratification by racial-ethnic group. Among whites, distant versus localized/regional stage did not influence the risk of being lost to follow-up. However, among Hispanics, APIs, and blacks, advanced stage cases were more likely to be lost to follow-up. Random censoring is a basic assumption of survival analysis (20), but these results show that censoring of SEER data is clearly not random across race-ethnicity. Because minorities were more likely to be lost to follow-up when diagnosed with late stage disease, censoring for non-whites was clearly associated with the outcome of interest, survival. In other words, non-whites diagnosed at late stage disease were more likely to die and were also more likely to be lost to follow up. This association introduces biases that produce an overestimation of survival among minority groups compared with whites. This suggests that the efficacy in assessing vital status differs across racial-ethnic group.

More complete death ascertainment is essential for unbiased survival comparisons. However, our analysis showed that a higher proportion of Hispanic and API cases especially if diagnosed with late stage disease were lost to follow-up compared with whites, which suggests a problem with missed deaths in cancer surveillance. There

### Table 1. Five-year survival by cancer site based on passive and active follow-up, SEER-18, 2000–2008* †

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>Site</th>
<th>Race/Ethnicity</th>
<th>Active follow-up (reported alive method)</th>
<th>Passive follow-up (presumed alive method)</th>
<th>Survival difference (in percent points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common cancers</td>
<td>Colorectal</td>
<td>Whites</td>
<td>56.0 (55.7–56.2)</td>
<td>56.2 (56.0–56.5)</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blacks</td>
<td>46.2 (45.6–46.8)</td>
<td>46.7 (46.1–47.3)</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hispanics</td>
<td>53.6 (52.9–54.4)</td>
<td>54.8 (54.1–55.5)</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APIs</td>
<td>59.8 (59.0–60.5)</td>
<td>60.9 (60.1–61.6)</td>
<td>1.1</td>
</tr>
<tr>
<td>Female breast</td>
<td>Whites</td>
<td>78.5 (78.3–78.6)</td>
<td>78.7 (78.5–78.8)</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>65.8 (65.1–66.4)</td>
<td>66.1 (65.5–66.7)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>75.9 (75.2–76.6)</td>
<td>76.8 (76.1–77.4)</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>81.9 (81.2–82.6)</td>
<td>82.4 (81.7–83.8)</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Lung and bronchus</td>
<td>Whites</td>
<td>14.7 (14.5–14.9)</td>
<td>14.9 (14.7–15.1)</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>10.6 (10.2–10.9)</td>
<td>10.9 (10.5–11.3)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>14.2 (13.6–14.8)</td>
<td>16.1 (15.6–16.8)</td>
<td>1.9†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>16.0 (15.3–16.6)</td>
<td>18.1 (17.4–18.7)</td>
<td>2.1†</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>Whites</td>
<td>82.4 (82.3–82.6)</td>
<td>82.5 (82.4–82.7)</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>75.2 (74.8–75.6)</td>
<td>75.5 (75.2–75.9)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>81.7 (81.2–82.2)</td>
<td>82.4 (82.0–82.9)</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>84.9 (84.4–85.5)</td>
<td>85.3 (84.8–85.8)</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Less favorable prognoses</td>
<td>Esophagus</td>
<td>Whites</td>
<td>15.4 (14.8–16.0)</td>
<td>15.6 (15.0–16.2)</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>8.8 (7.7–10.0)</td>
<td>9.1 (8.0–10.3)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>12.7 (10.9–14.5)</td>
<td>14.8 (13.0–16.7)</td>
<td>2.1†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>13.9 (11.6–16.5)</td>
<td>16.2 (13.8–18.8)</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Gall bladder</td>
<td>Whites</td>
<td>15.2 (13.8–16.7)</td>
<td>15.8 (14.4–17.3)</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>10.7 (8.3–13.4)</td>
<td>11.6 (9.2–14.4)</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>15.2 (13.0–17.4)</td>
<td>21.4 (19.2–23.7)</td>
<td>6.2†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>16.8 (13.6–20.4)</td>
<td>19.5 (16.2–23.0)</td>
<td>2.7†</td>
<td></td>
</tr>
<tr>
<td>Liver and intrahepatic bile duct</td>
<td>Whites</td>
<td>11.4 (10.9–11.9)</td>
<td>11.7 (11.2–12.2)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>7.1 (6.2–8.0)</td>
<td>8.0 (7.0–9.0)</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>9.6 (8.8–10.5)</td>
<td>11.6 (10.7–12.5)</td>
<td>2.0†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>16.5 (15.5–17.5)</td>
<td>18.4 (17.5–19.5)</td>
<td>1.9†</td>
<td></td>
</tr>
<tr>
<td>Pancreas</td>
<td>Whites</td>
<td>5.7 (5.5–6.0)</td>
<td>6.0 (5.7–6.3)</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>4.7 (4.2–5.3)</td>
<td>5.3 (4.7–5.9)</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>5.9 (5.2–6.6)</td>
<td>8.6 (7.8–9.4)</td>
<td>2.7†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>6.7 (5.8–7.7)</td>
<td>8.9 (8.0–10.0)</td>
<td>2.2†</td>
<td></td>
</tr>
<tr>
<td>Stomach</td>
<td>Whites</td>
<td>21.6 (21.1–22.2)</td>
<td>22.1 (21.5–22.7)</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blacks</td>
<td>21.2 (20.1–22.3)</td>
<td>21.7 (20.6–22.8)</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hispanics</td>
<td>21.5 (20.4–22.6)</td>
<td>24.9 (23.8–26.0)</td>
<td>3.4†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APIs</td>
<td>31.2 (29.9–32.5)</td>
<td>33.3 (32.0–34.5)</td>
<td>2.1</td>
<td></td>
</tr>
</tbody>
</table>

* All racial groups (whites, blacks, and Asian and Pacific Islanders [APIs]) exclude Hispanics. SEER = Surveillance, Epidemiology, and End Results.
† Observed, age-adjusted survival.
P < .05.
are many factors that can lead to missed deaths during linkage to death records, each of which can bias survival estimates (3,21,22). Finding a death match for cancer cases relies on variables such as date of birth, surname, and critically—a valid Social Security number (SSN) (23). Populations with large proportions of foreign-born subjects, such as APIs and Hispanics (and increasingly blacks), can be problematic for death linkage matches. This is primarily due to inaccurate or missing SSNs (3) or surname transcription errors. Currently, 13% of the US population is foreign born (16). The proportions of foreign-born Hispanics and APIs are much higher, 38% and 66%, respectively (24). In older age groups, which are most affected by cancer, these proportions are even higher. It is estimated that more than 10 million people use false SSNs in the United States (25), and the prevalence of invalid or absent SSNs is high among Hispanics and APIs (26). In regards to surnames, problems in identifying Hispanics and Asians include the order that names are presented, which differ from the Western convention. Often, paternal surnames occur before maternal surnames among...
Our findings stress the central aspect of a successful death linkage, which is the availability of a valid SSN. The groups with the lowest proportions of missed deaths were most likely to be either US nationals or US-born. These cases included not only whites and blacks but also specific Hispanic (Puerto Ricans and Cubans) and API subgroups (Japanese and Hawaiian). Puerto Ricans and Hawaiians are US citizens with valid SSNs available to them. The Cuban Adjustment Act of 1966 allows Cuban natives or nationals to become permanent residents after 1 year in the United States, making them a unique foreign-born population with access to legal status, a smooth naturalization process, and therefore a valid SSN. That is the likely reason for the relatively low proportion of missed deaths in this population. Japanese Americans are more likely to have a valid SSN—they are mostly US-born and Japanese immigrants do not follow the pattern of economic opportunity sought by other Asian and Hispanic immigrants. This can be attributed to a strong economy and high standard of living in Japan, especially since the second half of the 20th century. However, for other foreign-born groups, for example, Mexican, Filipino, Central and South American, legal immigration into the United States is more difficult and the inevitable proportion of undocumented immigrants is higher, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.

Moreover, causes for incomplete death linkages (e.g., inaccurate SSNs, mistyped name, inaccurate date of birth) are artificial in nature and independent of cancer stage or site. Thus, the proportion of missed deaths as a function of all deaths is likely to extend beyond highly fatal cancers to other stages and cancer sites. These differences in missed deaths will inflate survival estimates for Hispanics and APIs, not only for highly fatal cancers but also for all cancer stages and sites even though for cancer sites and stages with small numbers of deaths the effect on survival statistics may be negligible.

Although studies of cancer survival restricted to US-born Hispanics and US-born APIs are theoretically possible, with complete data on mortality and limited linkage-related bias, in practice they are not a valid possibility. This is because data regarding birthplace in most cancer registry datasets is often collected from death certificates, which is a biased data source in relation to survival. The proportion of cases with unknown birthplace is especially high, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.

Moreover, causes for incomplete death linkages (e.g., inaccurate SSNs, mistyped name, inaccurate date of birth) are artificial in nature and independent of cancer stage or site. Thus, the proportion of missed deaths as a function of all deaths is likely to extend beyond highly fatal cancers to other stages and cancer sites. These differences in missed deaths will inflate survival estimates for Hispanics and APIs, not only for highly fatal cancers but also for all cancer stages and sites even though for cancer sites and stages with small numbers of deaths the effect on survival statistics may be negligible.

Although studies of cancer survival restricted to US-born Hispanics and US-born APIs are theoretically possible, with complete data on mortality and limited linkage-related bias, in practice they are not a valid possibility. This is because data regarding birthplace in most cancer registry datasets is often collected from death certificates, which is a biased data source in relation to survival. The proportion of cases with unknown birthplace is especially high, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.

Moreover, causes for incomplete death linkages (e.g., inaccurate SSNs, mistyped name, inaccurate date of birth) are artificial in nature and independent of cancer stage or site. Thus, the proportion of missed deaths as a function of all deaths is likely to extend beyond highly fatal cancers to other stages and cancer sites. These differences in missed deaths will inflate survival estimates for Hispanics and APIs, not only for highly fatal cancers but also for all cancer stages and sites even though for cancer sites and stages with small numbers of deaths the effect on survival statistics may be negligible.

Although studies of cancer survival restricted to US-born Hispanics and US-born APIs are theoretically possible, with complete data on mortality and limited linkage-related bias, in practice they are not a valid possibility. This is because data regarding birthplace in most cancer registry datasets is often collected from death certificates, which is a biased data source in relation to survival. The proportion of cases with unknown birthplace is especially high, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.

Moreover, causes for incomplete death linkages (e.g., inaccurate SSNs, mistyped name, inaccurate date of birth) are artificial in nature and independent of cancer stage or site. Thus, the proportion of missed deaths as a function of all deaths is likely to extend beyond highly fatal cancers to other stages and cancer sites. These differences in missed deaths will inflate survival estimates for Hispanics and APIs, not only for highly fatal cancers but also for all cancer stages and sites even though for cancer sites and stages with small numbers of deaths the effect on survival statistics may be negligible.

Although studies of cancer survival restricted to US-born Hispanics and US-born APIs are theoretically possible, with complete data on mortality and limited linkage-related bias, in practice they are not a valid possibility. This is because data regarding birthplace in most cancer registry datasets is often collected from death certificates, which is a biased data source in relation to survival. The proportion of cases with unknown birthplace is especially high, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.

Moreover, causes for incomplete death linkages (e.g., inaccurate SSNs, mistyped name, inaccurate date of birth) are artificial in nature and independent of cancer stage or site. Thus, the proportion of missed deaths as a function of all deaths is likely to extend beyond highly fatal cancers to other stages and cancer sites. These differences in missed deaths will inflate survival estimates for Hispanics and APIs, not only for highly fatal cancers but also for all cancer stages and sites even though for cancer sites and stages with small numbers of deaths the effect on survival statistics may be negligible.

Although studies of cancer survival restricted to US-born Hispanics and US-born APIs are theoretically possible, with complete data on mortality and limited linkage-related bias, in practice they are not a valid possibility. This is because data regarding birthplace in most cancer registry datasets is often collected from death certificates, which is a biased data source in relation to survival. The proportion of cases with unknown birthplace is especially high, as is the potential for having invalid or absent SSNs. Accordingly, the proportion of missed deaths is elevated for these population subgroups. Overall, our results strongly suggest an overestimation of survival for population groups in which the prevalence of valid SSN is lower.
of last alive contact during the month of diagnosis, that is, with less than 1 month of survival time, and no date of death. SEER considers these individuals as potentially lost to follow-up or impossible to link, and as such, they are excluded in SEER default survival calculations. These cases cannot be identified in datasets that do not collect date of last contact (passive follow-up), so they are necessarily included in presumed alive survival and assumed to be alive with the maximum follow-up time. Thus, to simulate exactly the presumed alive assumption we included these cases in all analyses. However, in most situations, the reported alive survival statistics obtained by including or excluding cases with no survival times are often similar because of the very low proportion of cases with zero survival time in SEER, with the exception of cancers of very poor prognosis.

The inclusion or exclusion of these cases did not change the results of the lost to follow-up multivariate analyses in this study. Second, our comparisons between the reported alive and the presumed alive methods are likely to underestimate actual differences between the two types of follow-up. This is because SEER data were used in this study to simulate passive follow-up. The SEER follow-up system uses more sources of information to track cases and is likely to detect a higher number of deaths compared with the passive type of follow-up of non-SEER registries. Third, speculation that the proportion of deaths in distant stages extends to other stages discounts the salmon bias effect, which would be stronger for more advanced stages of cancer. However, other studies suggest that the actual effect of the salmon bias at the population basis is negligible (37).

In conclusion, our study found that passive follow-up and the presumed alive method inflate survival estimates for all racial-ethnic groups, but especially for Hispanics and APIs. Although active follow-up and SEER produce more accurate survival statistics for Hispanics and APIs, a bias of missed deaths in the death linkage persists and systematically inflates survival estimates for these two racial-ethnic groups. Across numerous racial-ethnic subgroups including Mexicans, South and Central Americans, Filipinos and South Asians, at least 4% of deaths are likely to be missed. These findings indicate that survival comparisons involving APIs and Hispanics can easily be biased. Although our findings cannot exclude a possible survival advantage for the foreign born or dismiss the potential for a healthy immigrant effect, researchers may want to consider residual differentials in death detection between racial-ethnic groups. Overestimation of survival and the absence of unbiased information on United States versus other place of birth can contribute to erroneous inferences regarding a presumed cancer survival advantage among Asians and Hispanics. Health advantages observed in Hispanics in comparison with non-Hispanic whites are commonly explained as part of the Hispanic Paradox, a term originally coined to characterize the lower Hispanic mortality despite their lower socio-economic status (32). However, as this study shows, the validity of such a paradox in cancer survival deserves further scrutiny. The problem of missed deaths makes accurate survival analysis for immigrant populations especially challenging. More complete follow-up of vital status in these populations is needed.

References


17. Mathew A. Removing Bias in Cancer Survival Estimates by Active Follow-up and Information on Determinants of Loss to Follow-up [dissertation]. Tampere, Finland: Acta Universitatis Tampereensis; 1996.


34. The Cuban Adjustment Act, 8 USC §1255 (1966).


**Funding**

National Institute of General Medical Sciences (8 P20 GM103440-11 to PSP).

**Affiliations of authors:** Epidemiology and Biostatistics, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV (PSP, TJB); California Cancer Reporting and Epidemiologic Surveillance (CalCARES) Program, Institute for Public Health Improvement, University of California Davis Health System, Sacramento, CA (CRM); Los Angeles Cancer Surveillance Program, Keck School of Medicine, University of Southern California, Los Angeles, CA (LL); Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Rockville, MD (SFA).