Cancer Causes & Control (2019) 30:21–29
https://doi.org/10.1007/s10552-018-1091-3

ORIGINAL PAPER

Concordance of cancer registry and self-reported race, ethnicity, and cancer type: a report from the American Cancer Society's studies of cancer survivors

Tracy M. Layne[1] · **Leah M. Ferrucci**[2] · **Beth A. Jones**[2] · **Tenbroeck Smith**[3] · **Lou Gonsalves**[4] · **Brenda Cartmel**[2]

Received: 7 December 2017 / Accepted: 23 October 2018 / Published online: 3 November 2018

Abstract

Purpose To examine the concordance between cancer registry and self-reported data for race, Hispanic ethnicity, and cancer type in the American Cancer Society's Studies of Cancer Survivors (SCS) I and II.

Methods We calculated sensitivity, specificity, positive predictive value, and Kappa statistics for SCS-I and II. The gold standard for cancer type was registry data and for race and ethnicity was self-reported questionnaire data.

Results Among 6,306 survivors in SCS-I and 9,170 in SCS-II, overall agreement (Kappa) for cancer type was 0.98 and 0.99, respectively. Concordance was strongest for breast and prostate cancer (Sensitivity ≥ 0.98 in SCS-I and II). For race, Kappa was 0.85 (SCS-I) and 0.93 (SCS-II), with strong concordance for white (Sensitivity = 0.95 in SCS-I and 0.99 in SCS-II) and black survivors (Sensitivity = 0.94 in SCS-I and 0.99 in SCS-II), but weak concordance for American Indian/Alaska Native (Sensitivity = 0.23 in SCS-I and 0.19 in SCS-II) and Asian/Pacific Islander survivors (Sensitivity = 0.43 in SCS-I and 0.87 in SCS-II). Agreement was moderate for Hispanic ethnicity (Kappa = 0.73 and 0.71; Sensitivity = 0.74 and 0.76, in SCS-I and SCS-II, respectively).

Conclusions We observed strong concordance between cancer registry data and self-report for cancer type in this national sample. For race and ethnicity, however, concordance varied significantly, with the poorest concordances observed for American Indian/Alaska Native and Asian/Pacific Islander survivors. Ensuring accurate recording of race/ethnicity data in registries is crucial for monitoring cancer trends and addressing cancer disparities among cancer survivors.

Keywords Cancer registries · Cancer survivors · Self-report · Disparities · Race · Ethnicity

---

✉ Tracy M. Layne
tracy.layne@nih.gov

Leah M. Ferrucci
leah.ferrucci@yale.edu

Beth A. Jones
beth.jones@yale.edu

Tenbroeck Smith
tenbroeck.smith@cancer.org

Lou Gonsalves
lou.gonsalves@ct.gov

Brenda Cartmel
brenda.cartmel@yale.edu

[1] Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, Room 6E324, MSC 9768, Bethesda, MD 20892, USA

[2] Department of Chronic Disease Epidemiology, Yale Comprehensive Cancer Center, Yale School of Public Health, 55 Church Street, Suite 801, New Haven, CT 06510, USA

[3] Behavioral and Epidemiology Research Group, American Cancer Society, 250 Williams Street, Atlanta, GA 30303, USA

[4] Connecticut Tumor Registry, Connecticut Department of Public Health, 410 Capitol Avenue, MS# 13TMR, Hartford, CT 06134-0308, USA

Introduction

Determining concordance between cancer registry data on race, ethnicity, and cancer diagnosis and corresponding self-reported data on these characteristics are useful for assessing the accuracy of each data source. Information from each source is used in different ways, including registry data to evaluate the cancer burden and trends in incidence and mortality by race and ethnicity [1] and self-reported cancer type to document self and/or family history of cancer [2–4]. Further, self-reported cancer history is relied upon by a number of large surveys, most notably national surveys, such as the National Health Interview Survey and the Behavioral Risk Factor Surveillance System, that are used to evaluate multiple domains of cancer survivorship [5].

Understanding and evaluating the validity of these data sources over time may be particularly relevant with changes in the demographics of the U.S. population and a growing body of research in disparities during cancer survivorship. With a rapidly growing cancer survivor population in the United States [6] and the projected increase in the number of racial and ethnic minority cancer patients [7], the accurate characterization of the survivorship experience within and across racial and ethnic groups is critical to improving the quality of life and addressing the long-term needs of these cancer survivors. Central cancer registries are an unparalleled resource for identifying individuals with cancer; however, accurate identification of minority populations requires accurate recording of the individual's race and Hispanic ethnicity.

Prior studies have examined the validity of self-report for cancer history [8–13] and cancer registries for race and ethnicity [14–17], although few used large national datasets and evaluated all races and ethnicities within the same study population. The gold standard for cancer type is the cancer registry, which is based on medical record data [18]. Much of the research in the U.S. evaluating cancer type concordance has been restricted to individual states and/or registries [8, 13, 19], with just one analysis examining racially/ethnically diverse groups across multiple non-Surveillance, Epidemiology, and End Results Program (SEER) registries [12]. Otherwise, prior research in this area has been mainly among women [9, 10] or restricted to a single cancer site [9]. Overall sensitivities for the concordance between registry data for cancer type and self-reported information range from 55 to 79% [8, 12, 13, 20, 21], with considerable variability across cancer sites in these and other studies of specific sub-types and cancer sites [9, 10]. In contrast to cancer type, self-reported information is the "gold standard" for capturing race and ethnicity and is the method that aligns with federal standards for reporting these characteristics [22, 23]. Self-reported race and ethnicity have been found to be inconsistent with data in cancer registries [16, 24]. Only one comprehensive study to date has examined consistency of race and ethnicity across multiple SEER registries [14], with other evaluations focused on individual states or registries [8, 13] or exclusively on a particular racial [25, 26] or ethnic group [27].

Here we report on the concordance between cancer registry and self-reported data with regards to cancer type, race, and Hispanic ethnicity in a large, geographically diverse population from state cancer registries, including SEER and non-SEER regions in the American Cancer Society's Studies of Cancer Survivors. This analysis adds to the current literature by covering a larger portion of the U.S. population than examined in previous studies on the concordance of cancer registry data and self-reported cancer type, race, and Hispanic ethnicity.

Methods

Study population

The American Cancer Society's Studies of Cancer Survivors explore quality of life in cancer survivors using a longitudinal design in the Study of Cancer Survivors-I (SCS-I) and a cross-sectional design in the Study of Cancer of Survivors-II (SCS-II) [28]. SCS-I enrolled survivors diagnosed between January 2000 and September 2003, approximately 1 year following diagnosis to assess predictors of adjustment to cancer and patterns of change in survivors. SCS-I survivors had one of the ten most common cancers: prostate, breast (female), lung, colorectal, bladder, non-Hodgkin's Lymphoma (NHL), melanoma, kidney, ovarian, or uterine cancer. SCS-II enrolled three groups of cancer survivors who were 2-, 5-, and 10-years post-diagnosis, with diagnoses occurring in 1999–2001, 1996–1998, and 1991–1993, respectively. To maximize the samples size in each of these groups, cancers with high (skin melanoma, breast, prostate, bladder, and uterine) or moderate (colorectal) 1- and 5-year survival rates were included [28].

Cancer survivors were identified through participating state cancer registries (i.e., 11 registries for SCS-I and 14 for SCS-II); registries were selected to be reflective of the U.S. population, and included states from each of the four Bureau of Census defined regions (West, Midwest, Northeast, and South) [28]. Minorities were oversampled in several states and survivors under age 55 were oversampled in SCS-I. In both studies, eligible survivors were 18 years or older at the time of diagnosis, had local, regional, or distant SEER Summary Stage cancer (bladder included in situ cases), spoke English or Spanish, and were residents of the state which had cancer registry coverage at the time of diagnosis. Survivors in both studies completed either self-administered

(95.5% and 89.1% in SCS-I and II, respectively) or telephone surveys in English or Spanish. Additional details have been described elsewhere [28]. Informed consent was obtained from all individual participants included in the study. Both SCS studies were approved by the Emory University Institutional Review Board, and by institutional review boards at participating registries, as required.

Measures

For self-reported cancer type, SCS-I respondents were asked at enrollment to "Please indicate which of the following types of cancer you have been diagnosed with during the past 2 years. Mark all that apply" Survivors selected from among the ten cancers listed above, as well as an option for "Other type of cancer (specify)." In SCS-II, survivors were asked "For each item, please mark if you have ever been diagnosed with that type of cancer and write down what year you were first diagnosed with that cancer." Survivors selected from among the six cancers listed above, in addition to an option for "Other (please specify)."

Respondents separately identified their race and ethnicity at enrollment. Survivors in both studies were asked "What describes your ethnic/racial background?," and could select from among the following options: African American/Black, American Indian/Alaska Native (AI/AN), Asian/Pacific Islander, Caucasian/White, and Other (Specify). In SCS-I, survivors could also select Multi-racial, and were told to "Mark all that apply," while in SCS-II, survivors had a "Multi-racial" option, and were asked to select only one option from the list. In both studies, survivors were also separately asked about ethnicity: "Do you consider yourself to be of Latino or Hispanic origin or descent?" to which they could respond "yes" or "no."

Data on cancer type, race, and Hispanic ethnicity were obtained directly from the state cancer registries. Data on a patient's race and ethnicity recorded in a cancer registry are based on medical record abstraction. These data include both self-reported and perceived race and ethnicity captured by hospital administrative staff or care providers, and/or inferred from last names (e.g., Asian race [29] and Hispanic ethnicity [17]). Basic demographic data were obtained via the self-reported questionnaire.

Analytical sample

In SCS-I, from the 6,306 available records, the cancer type analysis included 5,838 survivors after excluding those missing questionnaire cancer type ($n = 468$). The race analysis included 5,998 survivors following exclusion of missing self-reported race ($n = 176$), as well as reported being multiracial ($n = 60$), of some other race ($n = 58$), or missing registry race ($n = 1$). For Hispanic ethnicity, we excluded those with missing self-reported ethnicity only ($n = 518$), missing registry-reported ethnicity only ($n = 979$), and those missing both ($n = 26$), leaving a total of 4,780 survivors.

In SCS-II, from the 9,170 records, we excluded people with registry listed Non-Hodgkin Lymphoma ($n = 65$) as this was not listed as an individual cancer type in the self-report questionnaire and those with missing cancer type in the questionnaire ($n = 34$), leaving 9,071 survivors for the cancer type analysis. The race analysis excluded those with missing questionnaire race ($n = 386$), questionnaire "Other" race ($n = 216$), missing registry race ($n = 155$), questionnaire multiple races ($n = 125$), and those listed as "Other" by the registry ($n = 36$), leaving 8,255 survivors. As in SCS-I, we excluded survivors with missing data only in the registry ($n = 1{,}370$), questionnaire ($n = 643$), or both places ($n = 150$) for Hispanic ethnicity, resulting in a total of 7,007 survivors.

Statistical analysis

In SCS-I and SCS-II, measures of agreement between self-report and registry information, including sensitivity (Se), specificity (Sp), positive predictive value (PPV), and the Kappa (κ) statistic estimates and 95% confidence intervals (CI), were calculated separately for each cancer type and racial/ethnic category using Chi-square statistics. Additionally, the overall Kappa statistic was calculated for cancer type and race by looking at agreement for any of the individually examined cancer types and racial groups. An estimate of ≥ 80% was used as the cutoff for good/strong concordance for each measure [30]. All estimates of agreement were unchanged when the sample was restricted to survivors who completed self-administered, mail-in paper surveys, as such, final results include data from self-administered and telephone interview methods. All *p* values are two-sided and statistical analyses were conducted using SAS version 9.3 (SAS Institute, Cary, NC, USA).

Results

Survivor characteristics are presented in Table 1. The three state registries with the most individuals were Connecticut, New Jersey, and Minnesota in SCS-I, and Washington, Michigan, and California in SCS-II. Across SCS-I and II, estimates of the Se, Sp, PPV, and the Kappa for each cancer type were ≥ 0.94 for all measures (Table 2). In SCS-I and II, agreement estimates comparing survivors self-report to the gold standard registry-reported cancer were consistently high for breast (all estimates = 0.99) and prostate cancer (≥ 0.97 for all estimates). The overall Kappa statistics for cancer type were 0.98 in SCS-I and 0.99 in SCS-II (data not shown).

Table 1 Characteristics of survivors in the SCS studies, *n* (%)

Characteristic	SCS-I	SCS-II
No. of survivors	6,306	9,170
Age at questionnaire (Mean, S.D.)	59.9 (12.8)	67.3 (11.9)
Age at diagnosis (Mean, S.D.), years[a]	57.9 (19.5)	60.7 (11.9)
Time since diagnosis (Mean, S.D.), years	1.3 (0.33)	6.6 (3.3)
Cohort, years post-diagnosis		
2	–	3,229 (35.2)
5		3,270 (35.7)
10		2,671 (29.1)
Gender[a]		
Female	3,641 (57.7)	5,073 (55.3)
Male	2,664 (42.3)	4,097 (44.7)
Some college or more	3,394 (53.8)	2,865 (31.2)
Any health care coverage	4,942 (78.4)	8,933 (97.4)
Self-reported good health	2,181 (34.6)	3,573 (39.0)
State cancer registry		
Alabama	515 (8.2)	–
Arizona	–	370 (4.0)
California		1,316 (14.4)
Colorado		252 (2.8)
Connecticut	1,013 (16.1)	–
Delaware	–	401 (4.4)
Illinois		486 (5.3)
Indianapolis	228 (3.6)	–
Iowa	713 (11.3)	579 (6.3)
Maine	–	174 (1.9)
Massachusetts		795 (8.7)
Michigan		1,418 (15.5)
Minnesota	805 (12.8)	–
Nebraska	–	291 (3.2)
New Jersey	993 (15.8)	556 (6.1)
Ohio	522 (8.3)	–
Pennsylvania	684 (10.9)	405 (4.4)
Rhode Island	367 (5.8)	–
South Carolina	211 (3.40)	
Washington	–	1,778 (19.4)
Wyoming	255 (4.0)	349 (3.8)

[a] Per the registry

Estimates of Se, Sp, and PPV for race and ethnicity were wider in range across SCS-I and II (Table 3). Concordance was lowest AI/AN survivors (Se = 0.23 and 0.19 in SCS-I and II, respectively) and the highest for white survivors (as measured by Se = 0.95 in SCS-I and 0.99 in SCS-II). The overall Kappa statistics for race were 0.85 and 0.93 in SCS-I and II, respectively (data not shown). For Hispanic ethnicity, estimates ranged between 0.71 and 0.76 across SCS-I and II for all measures except Sp, which was 0.99 in SCS-I and 0.97 in SCS-II. The Kappa statistics for Hispanic ethnicity were 0.73 in SCS-I and 0.71 in SCS-II, respectively.

Discussion

This study examined the concordance between cancer registry data and self-report data on cancer type, race, and Hispanic ethnicity in a large national sample of state cancer registries and participants in the Studies of Cancer Survivors. We found strong agreement between self-report and the cancer registry for cancer type. Our finding of high concordance for all cancer types is consistent with other studies showing high sensitivity for cancers of the breast, prostate, and lung. This finding, however, runs contrary to reports of lower sensitivity documented for uterine (endometrial) cancer (e.g., 0.15 in EPIC-Spain and 0.69 in California Teachers Study) [10, 21], for reasons that are unclear. Researchers have also found strong levels of agreement in analyses of self-reported and registry linked data for breast, prostate, and large bowel cancers (Kappa: 0.81, 0.77, and 0.74, respectively) [8]. Research has also been undertaken among populations outside of the US, including the Netherlands [20], France [31], and Shanghai based on self-report and medical records data,[11] as well as registry data from Australia [32] and Spain [21]; with similar findings of greater sensitivity for breast cancers [32]. The accuracy of self-reported cancer type is critical for research aimed at characterizing the survivorship experience for specific cancers—this includes quality of life assessments, care coordination opportunities, identifying long-term care needs, and facilitating research on risk for subsequent cancers and/or other chronic conditions [5, 33]. This is also relevant as multiple national health surveys include self-reported cancer type (e.g., National Health Interview Survey [34], National Health and Nutrition Examination Survey [35], and the Behavioral Risk Factor Surveillance System [36]) and many researchers conduct survivorship analyses in these subsets of the survey populations.

For race and ethnicity, information was least accurate when comparing registry data to the self-report standard for AI/AN, Asian/Pacific Islanders, and Hispanics. Our findings for race and ethnicity are consistent with previous studies. While concordance between self-reported and registry data for race for non-Hispanic white and black race tends to be relatively high [14, 16, 25], estimates are consistently lower for AI/AN who often get misclassified as non-Hispanic white [14, 16, 26, 37, 38], Asian/Pacific Islanders [39, 40], and Latinos/Hispanics [14, 16, 27]. Moderate misclassification of Asian race (with some variability across sub-groups) and Hispanic ethnicity was observed in an evaluation of the Greater Bay Area Cancer Registry [16]. Race/ethnicity data housed in systematic data repositories, such as cancer registries, may represent information collected directly from included individuals. However, this information, at least historically, was often ascertained based on perceived race/ethnicity documented in the medical record by hospital

Table 2 Sensitivity, specificity, positive predictive value, and kappa statistic estimates for cancer type

	Registry-reported positive[a]	Self-reported positive	Sensitivity	Exact 95% CI	Registry-reported negative [a]	Self-reported negative	Specificity	Exact 95% CI	PPV	Exact 95% CI	Kappa	95% CI
SCS-I ($n=5{,}838$)												
Breast*	1,305	1,299	0.99	0.99–0.99	1,753	1,746	0.99	0.99–0.99	0.99	0.99–0.99	0.99	0.99–0.99
Prostate	991	980	0.98	0.98–0.99	1,193	1,183	0.99	0.98–0.99	0.99	0.98–0.99	0.98	0.97–0.98
Colorectal	763	750	0.98	0.97–0.99	4,480	4,469	0.99	0.99–0.99	0.99	0.97–0.99	0.98	0.97–0.99
Bladder	209	200	0.96	0.92–0.98	5,034	5,027	0.99	0.99–0.99	0.97	0.93–0.99	0.96	0.94–0.98
Uterine	276	259	0.94	0.90–0.96	2,782	2,779	0.99	0.99–0.99	0.99	0.97–0.99	0.96	0.94–0.98
Melanoma	290	285	0.98	0.96–0.99	4,953	4,946	0.99	0.99–0.99	0.98	0.95–0.99	0.98	0.97–0.99
NHL	332	327	0.98	0.97–0.99	4,911	4,904	0.99	0.99–0.99	0.98	0.96–0.99	0.98	0.97–0.99
Kidney	291	287	0.99	0.97–0.99	4,952	4,948	0.99	0.99–0.99	0.99	0.97–0.99	0.99	0.98–0.99
Lung	481	473	0.98	0.97–0.99	4,762	4,757	0.99	0.99–0.99	0.99	0.98–0.99	0.99	0.98–0.99
Ovarian	304	300	0.99	0.97–0.99	2,754	2,737	0.99	0.99–0.99	0.95	0.92–0.97	0.96	0.95–0.98
SCS-II ($n=9{,}071$)												
Breast*	2,655	2,651	0.99	0.99–0.99	1,787	1,770	0.99	0.98–0.99	0.99	0.99–0.99	0.99	0.99–0.99
Prostate	1,951	1,932	0.99	0.98–0.99	1,521	1,497	0.98	0.98–0.99	0.99	0.98–0.99	0.97	0.97–0.98
Colorectal	1,598	1,574	0.99	0.98–0.99	6,316	6,298	0.99	0.99–0.99	0.99	0.98–0.99	0.98	0.98–0.99
Bladder	471	448	0.95	0.93–0.97	7,443	7,434	0.99	0.99–0.99	0.98	0.96–0.99	0.96	0.95–0.98
Uterine	570	558	0.98	0.96–0.99	3,872	3,871	0.99	0.99–1.00	0.99	0.99–1.00	0.99	0.98–0.99
Melanoma	669	660	0.99	0.97–0.99	7,245	7,234	0.99	0.99–0.99	0.98	0.97–0.99	0.98	0.98–0.99

PPV positive predictive value

*Female breast cancer

[a] Gold standard

Table 3 Sensitivity, specificity, positive predictive value, and kappa statistic estimates for race and ethnicity

	Self-reported positive[a]	Registry positive	Sensitivity	Exact 95% CI	Registry negative	Self-reported negative[a]	Specificity	Exact 95% CI	PPV	Exact 95% CI	Kappa	95% CI
SCS-I (*n* = 5,998)												
American Indian/Alaska Native	22	5	0.23	0.078–0.45	5,976	5,974	0.99	0.99–1.00	0.71	0.29–0.96	0.34	0.12–0.57
Asian/Pacific Islander	35	15	0.43	0.26–0.61	5,963	5,882	0.99	0.98–0.99	0.16	0.09–0.24	0.22	0.13–0.32
Black	531	500	0.94	0.92–0.96	5,467	5,450	0.99	0.99–0.99	0.97	0.95–0.98	0.95	0.94–0.96
White	5,410	5,148	0.95	0.95–0.96	588	529	0.90	0.87–0.92	0.99	0.99–0.99	0.74	0.71–0.76
Hispanic[b]	206	152	0.74	0.67–0.80	3,480	3,431	0.99	0.98–0.99	0.76	0.69–0.81	0.73	0.68–0.78
SCS-II (8,255)												
American Indian/Alaska Native	58	11	0.19	0.10–0.31	8,197	8,192	0.99	0.99–0.99	0.69	0.41–0.89	0.30	0.16–0.43
Asian/Pacific Islander	212	185	0.87	0.82–0.91	8,043	8,024	0.99	0.99–0.99	0.91	0.86–0.94	0.89	0.85–0.92
Black	942	928	0.99	0.98–0.99	7,313	7,275	0.99	0.99–0.99	0.96	0.95–0.97	0.97	0.96–0.98
White	7,043	6,990	0.99	0.99–0.99	1,212	1,133	0.93	0.92–0.95	0.99	0.99–0.99	0.94	0.93–0.95
Hispanic[c]	630	480	0.76	0.73–0.79	6,377	6,181	0.97	0.96–0.97	0.71	0.67–0.74	0.71	0.68–0.73

PPV positive predictive value

[a] Gold standard

[b] *n* = 4,780 survivors

[c] *n* = 7,007 survivors

administrative staff or care providers, and/or inferred from last names (e.g., Asian race [29] and Hispanic ethnicity [17]). More recently, collection of race, ethnicity, and language preference have been included in the "meaningful use" objectives for electronic medical records, an initiative aimed at improving healthcare [41]. Despite recognition of these data as critical to improving care [42] and reducing racial/ethnic health disparities [43], it is not yet standard practice across healthcare institutions and there is no specification that self-reported race/ethnicity be used [42]. Of note, the cancer registry data used in the current study predate multiple initiatives to improve documentation of self-reported race and Hispanic ethnicity. To improve the accuracy of these data, cancer registries now undertake multiple activities, including annual linkage with Indian Health Service data to address the misclassification of American Indians and Alaska Natives [38]. In addition, cancer registries use algorithms to enhance the identification of Hispanic and Asian/Pacific Islander populations [44, 45].

Understanding the accuracy of race and ethnicity data elements in our data repositories is relevant as the racial and ethnic composition of the U.S. changes, with the fastest population growth expected among multi-racial individuals who are projected to triple in size by 2060 [46]. Additionally, our current racial/ethnic group categories are limited [47], combining heterogeneous sub-groups with varying cancer risks, as seen with individuals classified as Asian [48]. If researchers had the ability to disaggregate these data, there would be greater opportunity to better characterize incidence and mortality trends, as well as the cancer survivorship experience and address the needs across sub-groups that are underrepresented in research on survivors. This is particularly true for those with smaller population numbers and a higher burden of disease, such as Native Hawaiians and Pacific Islanders [48].

Our study leveraged data from a large national sample of cancer survivors from SEER and non-SEER registries, but was not without limitations. Several respondents were excluded due to missing self-reported data, though we had data on at least 75% of respondents for cancer type, race, and ethnicity in both studies. Additionally, relatively small sample sizes for specific racial/ethnic groups may have contributed to imprecise estimates of concordance statistics, as seen with results for Asian/Pacific Island survivors in SCS-I compared to SCS-II.

In summary, our study suggests that self-reported cancer type may be suitable for cancer survivorship research when registry data are not available. However, documentation of race and ethnicity can be improved, in part, through standardized collection, starting at the level of hospitals and health systems, based on self-reported race and ethnicity to best understand cancer trends across racial/ethnic groups. Inaccuracies related to these characteristics could impede research and exacerbate existing challenges to addressing racial and ethnic cancer disparities.

Acknowledgments We wish to acknowledge the cooperation and efforts of the cancer registries and public health departments from the states of Alabama, Arizona, California (Regions 2–6), Colorado, Connecticut, Delaware, Illinois, Iowa, Idaho, Maine, Massachusetts, Michigan, Minnesota, Nebraska, New Jersey, Ohio, Pennsylvania, Rhode Island, South Carolina, Washington, and Wyoming. We also thank the staff of the hundreds of hospitals, including Stamford Hospital, which reported cases to the participating cancer registries. Certain data used in this study were obtained from the Connecticut Tumor Registry located in the Connecticut Department of Public Health. Lastly, we are grateful to the thousands of cancer survivors, their physicians, and their loved ones who contributed to the collection of these data. The authors assume full responsibility for analyses and interpretation of these data.

Funding The American Cancer Society (ACS) Studies of Cancer Survivors (SCS) was supported by the intramural program of research conducted by the ACS Behavioral Research Center.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

References

1. Siegel RL, Miller KD, Jemal A (2016) Cancer statistics. CA Cancer J Clin 66(1):7–30. https://doi.org/10.3322/caac.21332
2. Kerber RA, Slattery ML (1997) Comparison of self-reported and database-linked family history of cancer data in a case-control study. Am J Epidemiol 146(3):244–248
3. Zajacova A, Dowd JB, Schoeni RF, Wallace RB (2010) Consistency and precision of cancer reporting in a multiwave national panel survey. Popul Health Metr 8:20. https://doi.org/10.1186/1478-7954-8-20
4. Murff HJ, Spigel DR, Syngal S (2004) Does this patient have a family history of cancer? An evidence-based analysis of the accuracy of family cancer history. JAMA 292(12):1480–1489. https://doi.org/10.1001/jama.292.12.1480
5. Lerro CC, Stein KD, Smith T, Virgo KS (2012) A systematic review of large-scale surveys of cancer survivors conducted in North America, 2000–2011. J Cancer Surviv 6(2):115–145. https://doi.org/10.1007/s11764-012-0214-1
6. de Moor JS, Mariotto AB, Parry C, Alfano CM, Padgett L, Kent EE, Forsythe L, Scoppa S, Hachey M, Rowland JH (2013) Cancer survivors in the United States: prevalence across the survivorship trajectory and implications for care. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 22 (4):561–570. https://doi.org/10.1158/1055-9965.EPI-12-1356
7. Smith BD, Smith GL, Hurria A, Hortobagyi GN, Buchholz TA (2009) Future of cancer incidence in the United States: burdens upon an aging, changing nation. J Clin Oncol 27(17):2758–2765. https://doi.org/10.1200/jco.2008.20.8983

8. Klein BE, Lee KE, Moss SE, Trentham-Dietz A, Klein R (2010) Self- and registry-reported cancer in a population-based longitudinal study. WMJ 109(5):261–266
9. Abraham L, Geller BM, Yankaskas BC, Bowles EJ, Karliner LS, Taplin SH, Miglioretti DL (2009) Accuracy of self-reported breast cancer among women undergoing mammography. Breast Cancer Res Treat 118(3):583–592. https://doi.org/10.1007/s10549-009-0375-4
10. Parikh-Patel A, Allen M, Wright WE, California Teachers Study Steering C (2003) Validation of self-reported cancers in the California Teachers Study. Am J Epidemiol 157(6):539–545
11. Gupta V, Gu K, Chen Z, Lu W, Shu XO, Zheng Y (2011) Concordance of self-reported and medical chart information on cancer diagnosis and treatment. BMC Med Res Methodol 11:72. https://doi.org/10.1186/1471-2288-11-72
12. Bergmann MM, Calle EE, Mervis CA, Miracle-McMahill HL, Thun MJ, Heath CW (1998) Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. Am J Epidemiol 147(6):556–562
13. Desai MM, Bruce ML, Desai RA, Druss BG (2001) Validity of self-reported cancer history: a comparison of health interview data and cancer registry records. Am J Epidemiol 153(3):299–306
14. Clegg LX, Reichman ME, Hankey BF, Miller BA, Lin YD, Johnson NJ, Schwartz SM, Bernstein L, Chen VW, Goodman MT, Gomez SL, Graff JJ, Lynch CF, Lin CC, Edwards BK (2007) Quality of race, Hispanic ethnicity, and immigrant status in population-based cancer registry data: implications for health disparity studies. Cancer Causes Control 18(2):177–187. https://doi.org/10.1007/s10552-006-0089-4
15. Hampton JW (2012) Network for cancer control research in American Indians and Alaska Natives: a historical perspective. J Cancer Educ 27(1 Suppl):S4–S8. https://doi.org/10.1007/s13187-012-0335-2
16. Gomez SL, Glaser SL (2006) Misclassification of race/ethnicity in a population-based cancer registry (United States). Cancer Causes Control 17(6):771–781. https://doi.org/10.1007/s10552-006-0013-y
17. Swallen KC, West DW, Stewart SL, Glaser SL, Horn-Ross PL (1997) Predictors of misclassification of Hispanic ethnicity in a population-based cancer registry. Ann Epidemiol 7(3):200–206
18. Ryerson AB, Eheman CR, Altekruse SF, Ward JW, Jemal A, Sherman RL, Henley SJ, Holtzman D, Lake A, Noone AM, Anderson RN, Ma J, Ly KN, Cronin KA, Penberthy L, Kohler BA (2016) Annual report to the nation on the status of Cancer, 1975–2012, featuring the increasing incidence of liver cancer. Cancer 122(9):1312–1337. https://doi.org/10.1002/cncr.29936
19. Mai PL, Garceau AO, Graubard BI, Dunn M, McNeel TS, Gonsalves L, Gail MH, Greene MH, Willis GB, Wideroff L (2011) Confirmation of family cancer history reported in a population-based survey. J Natl Cancer Inst 103(10):788–797. https://doi.org/10.1093/jnci/djr114
20. Schrijvers CT, Stronks K, van de Mheen DH, Coebergh JW, Mackenbach JP (1994) Validation of cancer prevalence data from a postal survey by comparison with cancer registry records. Am J Epidemiol 139(4):408–414
21. Navarro C, Chirlaque MD, Tormo MJ, Perez-Flores D, Rodriguez-Barranco M, Sanchez-Villegas A, Agudo A, Pera G, Amiano P, Dorronsoro M, Larranaga N, Quiros JR, Ardanaz E, Barricarte A, Martinez C, Sanchez MJ, Berenguer A, Gonzalez CA (2006) Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. J Epidemiol Community Health 60(7):593–599. https://doi.org/10.1136/jech.2005.039131
22. Sohn MW, Zhang H, Arnold N, Stroupe K, Taylor BC, Wilt TJ, Hynes DM (2006) Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. Popul Health Metr 4(7):7. https://doi.org/10.1186/1478-7954-4-7
23. Office of Management and Budget (1997) Revisions to the standard for the classification of federal data on race and ethnicity. Fed Reg 62(210):58782–58790
24. Frost F, Taylor V, Fries E (1992) Racial misclassification of Native Americans in a surveillance, epidemiology, and end results cancer registry. J Natl Cancer Inst 84(12):957–962
25. Patel DA, Knowles A, Schwartz AG, Schwartz K (2005) Evaluation of African-American and white racial classification in a surveillance, epidemiology, and end results cancer registry. Ethn Dis 15(4):713–719
26. Hoopes MJ, Taualii M, Weiser TM, Brucker R, Becker TM (2010) Including self-reported race to improve cancer surveillance data for American Indians and Alaska Natives in Washington state. J Reg Manag 37(2):43–48
27. Stewart SL, Swallen KC, Glaser SL, Horn-Ross PL, West DW (1999) Comparison of methods for classifying Hispanic ethnicity in a population-based cancer registry. Am J Epidemiol 149(11):1063–1071
28. Smith T, Stein KD, Mehta CC, Kaw C, Kepner JL, Buskirk T, Stafford J, Baker F (2007) The rationale, design, and implementation of the American Cancer Society's studies of cancer survivors. Cancer 109(1):1–12. https://doi.org/10.1002/cncr.22387
29. Hsieh MC, Pareti LA, Chen VW (2011) Using NAPIIA to improve the accuracy of Asian race codes in registry data. J Reg Manag 38(4):190–195
30. Watson PF, Petrie A (2010) Method agreement analysis: a review of correct methodology. Theriogenology 73(9):1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003
31. Berthier F, Grosclaude P, Bocquet H, Faliu B, Cayla F, Machelard-Roumagnac M (1997) Prevalence of cancer in the elderly: discrepancies between self reported and registry data. Br J Cancer 75(3):445–447. https://doi.org/10.1038/bjc.1997.74
32. Stavrou E, Vajdic CM, Loxton D, Pearson SA (2011) The validity of self-reported cancer diagnoses and factors associated with accurate reporting in a cohort of older Australian women. Cancer Epidemiol 35(6):e75–e80. https://doi.org/10.1016/j.canep.2011.02.005
33. Homan SG, Kayani N, Yun S (2016) Risk factors, preventive practices, and health care among breast cancer survivors, United States, 2010. Prev Chronic Dis 13:E09. https://doi.org/10.5888/pcd13.150377
34. Tarasenko Y, Chen C, Schoenberg N (2017) Self-reported physical activity levels of older cancer survivors: results from the 2014 national health interview survey. J Am Geriatr Soc 65(2):e39–e44. https://doi.org/10.1111/jgs.14589
35. Liu B, Taioli E (2014) Associations between human papillomavirus and history of cancer among U.S. adults in the National Health and Nutrition Examination Survey (2003–2010). Br J Cancer 111(7):1448–1453. https://doi.org/10.1038/bjc.2014.414
36. Tarleton HP, Ryan-Ibarra S, Induni M (2014) Chronic disease burden among cancer survivors in the California Behavioral Risk Factor Surveillance System, 2009–2010. J Cancer Surviv 8(3):448–459. https://doi.org/10.1007/s11764-014-0350-x
37. Johnson JC, Soliman AS, Tadgerson D, Copeland GE, Seefeld DA, Pingatore NL, Haverkate R, Banerjee M, Roubidoux MA (2009) Tribal linkage and race data quality for American Indians in a state cancer registry. Am J Prev Med 36(6):549–554. https://doi.org/10.1016/j.amepre.2009.01.035
38. Espey DK, Wiggins CL, Jim MA, Miller BA, Johnson CJ, Becker TM (2008) Methods for improving cancer surveillance

data in American Indian and Alaska Native populations. Cancer 113(5 Suppl):1120–1130. https://doi.org/10.1002/cncr.23724
39. Gomez SL, Glaser SL, Kelsey JL, Lee MM (2004) Bias in completeness of birthplace data for Asian groups in a population-based cancer registry (United States). Cancer Causes Control 15(3):243–253. https://doi.org/10.1023/b:caco.0000024244.91775.64
40. Swallen KC, Glaser SL, Stewart SL, West DW, Jenkins CN, McPhee SJ (1998) Accuracy of racial classification of Vietnamese patients in a population-based cancer registry. Ethn Dis 8(2):218–227
41. Blumenthal D, Tavenner M (2010) The “meaningful use” regulation for electronic health records. New Engl J Med 363(6):501–504. https://doi.org/10.1056/NEJMp1006114
42. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, Kontos EZ, Park ER, Marinacci LX, Haas JS (2015) Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med 30(6):719–723. https://doi.org/10.1007/s11606-014-3102-8
43. IOM (Institute of Medicine) (2009) Race, ethnicity, and language data: standardization for health care quality improvement. IOM (Institute of Medicine), Washington, DC
44. Group NRaEW (2011) NAACCR guideline for enhancing hispanic/latino identification: revised NAACCR Hispanic/Latino identification algorithm. vol Hispanic/Latino identification: revised NAACCR Hispanic/Latino identification algorithm [NHIA v2.2.1]. Springfield
45. Group NRaEW NAACCR Asian Pacific Islander Identification Algorithm [NAPIIA v1.2.1]. Springfield
46. Colby SL, Ortman JM (2014) Projections of the size and composition of the U.S. population: 2014–2060
47. Brawley OW (2016) Some thoughts on health surveillance data, race, and population categorization. CA Cancer J Clin 66(3):179–181. https://doi.org/10.3322/caac.21346
48. Torre LA, Sauer AM, Chen MS Jr, Kagawa-Singer M, Jemal A, Siegel RL (2016) Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: converging incidence in males and females. CA Cancer J Clin 66(3):182–202. https://doi.org/10.3322/caac.21335