HHS Public Access
Author manuscript
Cancer Causes Control. Author manuscript; available in PMC 2025 April 01.

Published in final edited form as:
Cancer Causes Control. 2024 April ; 35(4): 685–694. doi:10.1007/s10552-023-01827-3.

Accuracy of patient race and ethnicity data in a central cancer registry

Rachel R. Codden[1,2], **Carol Sweeney**[1,2,3], **Blessing S. Ofori-Atta**[1], **Kimberly A. Herget**[2], **Kacey Wigren**[2], **Sandra Edwards**[4], **Marjorie E. Carter**[2], **Rachel D. McCarty**[3,6], **Mia Hashibe**[2,3,5], **Jennifer A. Doherty**[2,3,6], **Morgan M. Millar**[1,2,3]

[1] Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA

[2] Utah Cancer Registry, University of Utah, Salt Lake City, UT, USA

[3] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

[4] Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA

[5] Department of Family and Preventive Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA

[6] Department of Population Health Sciences, University of Utah School of Medicine, Salt Lake City, UT, USA

Abstract

Purpose—Race and Hispanic ethnicity data can be challenging for central cancer registries to collect. We evaluated the accuracy of the race and Hispanic ethnicity variables collected by the Utah Cancer Registry compared to self-report.

Methods—Participants were 3,162 cancer survivors who completed questionnaires administered in 2015–2022 by the Utah Cancer Registry. Each survey included separate questions collecting race and Hispanic ethnicity, respectively. Registry-collected race and Hispanic ethnicity were

Corresponding author: Morgan M. Millar, Morgan.millar@hsc.utah.edu.

Author Contributions
R.R.C., M.M.M., C.S., M.E.C., and K.A.H. contributed to the study conception and design. M.M.M., J.A.D., R.D.M., and M.H. contributed to data acquisition. S.E. and M.E.C. oversaw material preparation and data collection. Analyses were performed by B.S.O.-A., K.A.H., and R.R.C. The first draft of the manuscript was written by R.R.C. K.W. revised the draft critically for important intellectual content. M.M.M. and C.S. assisted in substantive revisions to the drafted manuscript. All authors read and approved the final manuscript.

Competing Interests
The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval
The Utah Cancer Survivors Study (Study 1 in this analysis) was approved by the University of Utah Institutional Review Board (IRB no. 00065816). The University of Utah Institutional Review Board reviewed and granted exemption to the SEER Program Rapid Response Surveillance System study (Study 2, IRB no. 00096345) and the Malignancy Health and Lifestyle Study (Study 4, IRB no. 00127902). The Cancer Survivor Experiences Survey (Study 3) was reviewed by the Utah Department of Health and Human Services Institutional Review Board, which deemed the project exempt from human subjects research approval because it is a program evaluation initiative.

Consent to Participate
Informed consent was obtained from all participants included in this study.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

compared to self-reported values for the same individuals. We calculated sensitivity and specificity for each race category and Hispanic ethnicity separately.

Results—Survey participants included 323 (10.2%) survivors identifying as Hispanic, a lower proportion Hispanic than the 12.1% in the registry Hispanic variable (sensitivity 88.2%, specificity 96.5%). For race, 43 participants (1.4%) self-identified as American Indian or Alaska Native (AIAN), 32 (1.0%) as Asian, 23 (0.7%) as Black or African American, 16 (0.5%) Pacific Islander (PI), and 2994 (94.7%) as White. The registry race variable classified a smaller proportion of survivors as members of each of these race groups except White. Sensitivity for classification of race as AIAN was 9.3%, Asian 40.6%, Black 60.9%, PI 25.0%, and specificity for each of these groups was >99%. Sensitivity and specificity for White were 98.8% and 47.4%.

Conclusion—Cancer registry race and Hispanic ethnicity data often did not match the individual's self-identification. Of particular concern is the high proportion of AIAN individuals whose race is misclassified. Continued attention should be directed to the accurate capture of race and ethnicity data by hospitals.

Keywords

Race; Hispanic ethnicity; cancer disparities; cancer survivor; cancer registries; validity

Introduction

There is a growing body of evidence identifying cancer disparities in the United States by race and ethnicity. For example, from 2015-2019 the cancer mortality rate per 100,000 was highest in Black individuals at 179, followed by 161 in American Indian or Alaska Native (AIAN) individuals, 157 in White individuals, 110 in Hispanic or Latino individuals, and lowest in Asian or Pacific Islander (API) individuals at 96 [1]. AIAN populations have the highest incidence rates of liver and intrahepatic bile duct cancer, with Hispanic or Latino and API populations following next [2]. Another striking fact is although Black and White women have a similar incidence rate of breast cancer, Black women are more likely to die of the cancer [2, 3].

To fully understand cancer disparities, it is essential that cancer surveillance data, which are the foundation of monitoring population-based trends in cancer incidence and mortality, contain accurate reports of the race and ethnicity of the cancer patients. Patient race and Hispanic ethnicity are collected as standardized data items by central cancer registries in the United States [4]. Central registries rely primarily on cancer case abstracts submitted by hospitals for obtaining patient race and ethnicity data.

The gold-standard for race and ethnicity are self-reported values from the individual [5]. Several reports of the validity of cancer registry race and ethnicity data compared to self-report have been based on large study populations spanning multiple states and multiple cancer registries [6-8]. These studies found that accurate registry classification of race for individuals diagnosed with cancer was highly sensitive, 95-99%, for those of White race, and classification of Black or African American also high at 91-99%. Classification of individuals of Hispanic ethnicity was less accurate in these studies, with sensitivity of

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

74-79%. Classification of individuals diagnosed with cancer as Asian and Pacific Islander (these two race categories have historically been combined in cancer reporting) and AIAN have lower accuracy in the same studies. Classification of cancer patients as AIAN had notably poor sensitivity, 19-40%, with these estimates being based on fewer than 100 participants who self-reported race as AIAN in each study [6-8].

Self-report data in one of these multi-state studies was based on data collected in 2003-2011 [7]. The other two studies were based on cancer registry data collected in 2001 or earlier, with one study using self-report and cancer registry data collected as far back as 1973 [8]. The proportions of the U.S. population who are members of different race and ethnic groups has changed over time, including an increasing proportion who report more than one race [9, 10]. Health care systems may change their strategies for collecting race and ethnicity information about patients to meet evolving standards. These factors will in turn cause the quality of race and ethnicity data collected by cancer registries to change over time, thus validity should continue to be evaluated using recent data.

The practice of grouping all individuals of Asian or Pacific Islander origin into a single category when reporting cancer data has been criticized, especially considering the different cancer incidence patterns observed for Asian compared to Pacific Islander populations in the United States [11, 12]. Thus, it is important to assess the validity of race classification in more specific subgroups. To our knowledge, only two studies, both using study populations in the state of California, have assessed validity of cancer registry data for classifying subgroups defined by national origin within the broad category of Asian or Pacific Islander [13, 14].

A limited number of studies have reported validity of race and/or ethnicity for central cancer registries within single states [8, 13-20]. Results from these single-state studies are suggestive that accuracy may vary by geographic region. For example, reports of sensitivity of classification of ethnicity as Hispanic range from as high as 98% in Michigan [17] and 88% in Illinois [18], to lower sensitivities of 66-70% in California [14, 15, 20] and 58% previously in Utah [16].

Investigating the accuracy of race and ethnicity data describing residents of the state of Utah who are diagnosed with cancer is important as the Utah population is expanding and diversifying rapidly. According to Census data, Utah was the fastest-growing state between 2010 and 2020 [21]. The population increased 18.4% over that 10-year span, with a 7.1 percentage point increase in its diversity index (33.6% in 2010 vs. 40.7% in 2020), which measures the probability that two people chosen at random in a given area will be from two different race or ethnic groups [21]. Furthermore, investigating accuracy of the registry’s identification of AIAN race is of special interest because we have reported that the overall cancer incidence in the Utah AIAN population appears to be dramatically lower than other race groups [22]. The second-largest racial or ethnic group in Utah is the Hispanic or Latino population, representing 14.4% of the state population [23]. Therefore, it is important to understand if the North American Association of Central Cancer Registries’ Hispanic Identification Algorithm (NHIA), a tool to assist registries in capturing Hispanic ethnicity data, adds value to the cancer registry ethnicity classification. Also, the Pacific Islanders

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

population in Utah is growing and it is important to treat Asian and Pacific Islander as separate categories.

The purpose of this analysis was to evaluate the accuracy of the race and ethnicity variables as collected by the Utah Cancer Registry by comparing them to self-reported data obtained directly from cancer survivors. In this analysis we also examine predictors of accuracy in identification of AIAN race. Given the documented challenges for cancer registries nationally to obtain valid classification of AIAN race, it is possible that cancer incidence in this group in Utah may be underestimated due to misclassification of race. We also evaluate whether NHIA adds value to the cancer registry Hispanic ethnicity classification and evaluate validity of the categories of Asian and Pacific Islander separately. By evaluating the accuracy of Utah Cancer Registry's race and ethnicity variables, this study will inform cancer control efforts and cancer research that relies on registry data to understand the extent of disparities in cancer and address them.

Methods

Participants and data sources

The Utah Cancer Registry is a population-based central registry that collects and maintains information on all reportable cancer diagnoses in Utah. Utah Cancer Registry data are complete and of high quality according to the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program and the U.S. Centers for Disease Control and Prevention's National Program of Cancer Registries [24]. Race and ethnicity data are frequently reported to the registry through hospital-submitted cancer case abstracts. Less commonly these data are provided by radiation treatment centers, pathology laboratories, or physician offices. The North American Association of Central Cancer Registries (NAACCR) defines data standards for coding variables, including race codes to be used for up to five races per cancer patient and prioritization of race codes when an individual has multiple reported races. Hispanic ethnicity is a separate variable from the race variable [4].

The registry utilizes linkages and algorithms to fill in missing race and ethnicity information. Linkages to Utah death and birth certificates and Indian Health Services (IHS) enrollment records are performed. NHIA identifies individuals as Hispanic based on birth country, maiden name, and current surname and can be applied to individuals diagnosed in 1995 and later [25]. The NAACCR Asian Pacific Islander Identification Algorithm (NAPIIA) is similarly used by cancer registries to identify Asian/Pacific Islander individuals based on surname when race information is missing [26].

The validity of registry race and ethnicity data was assessed by comparing it with self-reported race and ethnicity data obtained through four recent surveys conducted for cancer research or public health projects and administered by the Utah Cancer Registry. The Utah Cancer Survivors Study (administered 2015, 2018; Study 1 in Table 1) [27] and the Cancer Survivors' Experiences Survey (administered 2019-2022; Study 3 in Table 1) [28] focused on health status and quality of life among survivors of cancer of multiple sites. Additional participants were included from a SEER Program Rapid Response Surveillance System study to determine feasibility of obtaining patient-reported outcomes from cancer survivors

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

(administered 2017; Study 2 in Table 1) [29]. Finally, participants in the Malignancy Health and Lifestyle Study (administered 2020-2022; Study 4 in Table 1) [30] were cases selected for case-control studies of melanoma, leukemia, and lymphoma. The aims, study design, eligibility criteria, and methods of these studies have been previously reported. The administration of these surveys followed Utah policy regarding surveillance data, which requires the initial contact with cancer patients for research recruitment be made by Utah Cancer Registry as the surveillance data steward. All studies were reviewed by the University of Utah or the Utah Department of Health and Human Services Institutional Review Board.

All individuals who responded to one of the four surveys and answered the race and/or ethnicity survey questions were included in analyses. The analysis did not include responses which were missing both the race and ethnicity data. For participants that were diagnosed with more than one cancer, the cancer characteristics were reported for the primary cancer that met the specific study eligibility. In the instance that an individual responded to more than one survey (n=18), the more recent survey response was used. The race and ethnicity data from the survey response were combined with data from the cancer patients' registry records for analysis.

Variables

The four surveys providing self-reported race and ethnicity data contained similarly structured questions requesting individuals to report their race and ethnicity. In three of the surveys (Studies 1, 3, 4 in Table 1), participants were asked whether they are of Hispanic, Latino/a, or Spanish origin with Yes and No response options. The other survey (Study 2 in Table 1) asked the participant whether their ethnicity is Hispanic or Non-Hispanic. One survey asked an additional question to identify the specific country of origin, however this information was not used in the analysis. All four surveys asked the participants to identify their race in a "Select all that apply" format. The race options in each survey were similarly listed as White, Black or African American, American Indian or Alaska Native, Asian, and Pacific Islander. Additional responses such as "Other" and "Don't know/Not sure" varied among the surveys.

For purposes of this analysis, we created a series of binary variables representing each self-reported race (Black or African American, Asian, Pacific Islander, American Indian or Alaska Native, and White) and a binary variable representing self-reported Hispanic ethnicity. Similarly, we created binary variables representing each registry race category and registry Hispanic ethnicity. For every participant, each of these variables was classified as Yes or No. In the event that a participant self-reported more than one race in the survey, Yes was indicated in multiple categories. Therefore, an individual participant may have more than one documented self-reported race represented in the analysis. There were no individuals in the analysis for which the registry had more than one race documented.

Additional demographic and cancer variables used in this analysis were obtained from registry records. These include sex, age at time of survey, cancer site, diagnosis year, and address of residence at diagnosis. The age at survey for Study 2, Study 3, and Study 4 was calculated as the difference between date of birth and date of survey completion. Age at

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

survey for Study 1 was estimated as the difference between year of survey completion and year of birth (Table 1). The addresses were classified as urban or rural at the county level according to USDA's Rural-Urban Continuum Codes [31].

Statistical analysis

We compiled a dataset of responses to the four surveys along with matched registry variables. Descriptive statistics (counts and percentages) were calculated for the entire sample and stratified by survey. Individuals who did not provide a response to the survey questions collecting self-reported Hispanic ethnicity (n=23) or race (n=71), are excluded from the remainder of our analyses as self-reported race data are necessary for calculation of the statistics described below.

We classified each participant as Yes vs. No for self-report and Yes vs. No for registry report for Hispanic ethnicity and for each race group separately. We then calculated sensitivity, specificity, and positive predictive values with 95% confidence intervals (CI) for Hispanic ethnicity and for each race. With this structure, each race group may contain individuals who identify as either Hispanic or non-Hispanic. Individuals who self-reported more than one race (n=37) are included in these calculations for each race they reported, and thus included in counts for each race they reported. Individuals who reported "Other" race (n=29) were included in sensitivity and specificity calculations for the remaining race groups, however we did not calculate sensitivity and specificity for the "Other" race category as it is a composite category representing multiple uniquely reported identities that cannot be directly compared to registry data.

We further summarized sensitivity, specificity, and positive predicitive values for AIAN race according to several other variables including Hispanic ethnicity, age at survey, and diagnosis year. Finally, we assessed the accuracy of the NHIA algorithm by calculating the percentage of individuals with unknown ethnicity in registry records who were correctly classified as Hispanic or non-Hispanic after implementation of the algorithm. All analyses were conducted with R 4.1.3 and SAS Version 9.4.

Results

After exclusion of incomplete registry records (n=68) and duplicate records for individuals participating in more than one of the surveys (n=18), our analysis comprised a total of 3,162 individuals diagnosed with cancer. Descriptive statistics for the total study sample and stratified by study are shown in Table 1. Males comprised 48.7% of the total sample and females 51.3%. A majority of participants were in the age range of 55-74 years (57.6%) at survey completion and resided in urban areas (87.3%).

Approximately half of the participants originated from Study 3 which surveyed cancer survivors of all cancer sites. Inclusion criteria for the other three studies were restricted to specific cancer sites, influencing the cancer sites represented in this analysis. The sites with highest counts of participants are melanoma (23.2%), lymphoma (13.0%), and breast cancer (11.7%). Three of the four studies limited eligibility to participants based on recency of cancer diagnosis, also influencing the distribution of cancer diagnosis year for this

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

analysis. The great majority of participants were recently diagnosed, with 36.8% diagnosed in 2019-2021. Only 12.1% of participants were diagnosed in 2012 or prior.

As shown in Table 2, 323 (10.2%) participants reported Hispanic ethnicity, a somewhat lower proportion Hispanic than the 12.1% indicated by the registry Hispanic variable. Forty-three participants (1.4%) identified as American Indian or Alaska Native (AIAN), 32 (1.0%) as Asian, 23 (0.7%) identified as Black or African American, and 16 (0.5%) as Pacific Islander (PI). The registry race variable indicated a smaller proportion of survivors to be members of these race groups: 0.3% AIAN, 0.6% Asian, and 0.5% Black or African American. Based on self-report, 94.7% of participants were White, similar to the proportion in registry records (95.2%).

Sensitivity and specificity of registry classification of Hispanic ethnicity were reasonably strong, 88.2% (95% CI 84.2, 91.5) and 96.5% (95% CI 95.8, 97.2), respectively. A sensitivity analysis was completed due to a variation in the wording of the ethnicity question in Survey 2. The results showed near identical sensitivity for Hispanic ethnicity when excluding Survey 2, thus it does not appear that the variation in survey wording had an impact on results. Sensitivity for survey participants self-reporting race as White was very high, 98.8% (95% CI 98.3, 99.2), but specificity was low at 47.4% (95% CI 37.2, 57.8). Sensitivity for registry classification of race for survey participants self-reporting in other groups was lower, with sensitivity for Black or African American 60.9% (95% CI 38.5, 80.3), Asian 40.6% (95% CI 23.7, 59.4), Pacific Islander 25.0% (95% CI 7.3, 52.4), and AIAN 9.3% (95% CI 2.6, 22.1). Specificity for each of these race groups was greater than 99%. The highest positive predictive value was for White participants (98.3%, 95% CI 97.8, 98.7), meaning that among those whom the registry records classified as White, the probability of also self-reporting as such was 98.3%. The next highest positive predictive value was for Black or African American participants (93.3%, 95% CI 68.1, 99.8), followed by Hispanic participants (74.4%, 95% CI 69.7, 78.7). The lowest positive predictive value was for AIAN participants, with a value of 36.4% (95% CI 10.9, 69.2).

We also conducted a supplemental analysis excluding the 37 individuals from our sample who self-reported more than one race (Supplemental Table 1). The sensitivity for races other than White had slight increases but remained low. These values were 21.1% (95% CI 6.1, 45.6) for AIAN participants, 57.1% (95% CI 34.0, 78.2) for Asian participants, 81.2% (95% CI 54.4, 96.0) for Black or African American participants, and 40.0% (95% CI 12.2, 73.8) for Pacific Islander participants.

To further evaluate accuracy for AIAN race, we examined sensitivity and specificity of AIAN classification stratified by additional variables including self-reported Hispanic ethnicity, age, and diagnosis year (Table 3). The number of observations in each category were small and confidence intervals were wide. The calculated sensitivity for AIAN race was low in both self-identifying Hispanic (4.3%, 95% CI 0.1, 21.9) and non-Hispanic (10.5%, 95% CI 1.3, 33.1) participants.

To assess the accuracy of the NHIA algorithm, which classifies cancer patients as Hispanic or non-Hispanic, we compared results of the algorithm to participants' self-report. There

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

were 40 participants whose Hispanic ethnicity status was unknown to the registry prior to utilization of the NHIA algorithm. For these individuals, self-reported Hispanic ethnicity matched the ethnicity categorization of the algorithm 100% of the time (results not shown).

Lastly, a supplemental analysis was completed using a combined race and ethnicity variable wherein each race was evaluated exclusive of Hispanic identity. These findings were substantively similar to our primary analysis (see Supplemental Table 2). We found that the sensitivity of non-Hispanic White was 95.5% (95% CI 94.7, 96.3), which is slightly lower than sensitivity for all White individuals. Non-Hispanic White specificity was 84.3% (95% CI 79.8, 88.1), which is higher than the overall White specificity. Sensitivity for non-Hispanic Black or African American was 68.8% (95% CI 41.3, 89.0), which is somewhat higher than when not limiting to non-Hispanic.

Discussion

This study found that the race information contained in Utah Cancer Registry records was very often valid for individuals identifying as White. Among those whom the registry classified as White, the probability of also self-reporting as White was 98.3%. Sensitivity for classification as Hispanic was reasonably high at 88.2%, a substantial improvement over the 58% sensitivity reported from our earlier evaluation based on females diagnosed in the 1990's [16]. However, sensitivity was substantially lower for every race group other than White. The most notable low sensitivity was 9.3% for individuals identifying as AIAN, meaning that among participants who self-identified as AIAN, 90.7% were coded in the registry to race categories other than AIAN. Because the proportion who report race as AIAN is higher among Hispanic than non-Hispanic individuals [32], we considered whether sensitivity of classification as AIAN might vary by Hispanic ethnicity. We found that validity of AIAN classification was poor in both Hispanic and non-Hispanic populations. Prior studies also found sensitivity of classification of AIAN by cancer registries to be poor [6-8, 14]. Our more recent data show that the situation is not improving, although the small sample size and large confidence intervals make it difficult to draw firm conclusions regarding what factors are associated with higher sensitivity for AIAN race.

While looking at the data on a case-by-case level, of those who self-reported a race other than White, a higher percentage of these individuals were incorrectly classified as White in registry records than were correctly identified. This is further evidenced by the specificity of 47.4% for White individuals, which is the lowest specificity seen in our results. This means that Utah Cancer Registry is undercounting cancer cases among AIAN, Asian, Black or African American, and Pacific Islander individuals, and that cancer incidence for these race groups in the state may be underestimated. This is further exacerbated by the results of our supplemental analysis which showed when individuals who self-reported more than one race were excluded, sensitivities for races other than White increased. From this we can see when individuals have only one self-reported race, the registry has more accurately captured their race, however when participants reported more than one race, the registry did not accurately reflect so. Registry variables allow registries to record up to five races per individual to better capture the identifies of individuals who self-identify as more than one

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

race. Our analyses confirm the importance of ascertaining whether registries are capturing patients' complete racial identities.

Previous research from multi-state studies and from other state cancer registries also found that sensitivity is high for White individuals. Other studies have reported higher sensitivity for Black race than what we observed in Utah, a state in which only 1.2% of the population is Black or African American [6, 8, 18]. Most other studies have treated Asian and Pacific Islander populations as a single group. Separating these two groups is particularly important as the Pacific Islander community is one of the fastest growing populations in Utah [21, 33]. We observed higher sensitivity for Asian than for Pacific Islander populations, 41% compared to 25%. Gomez and colleagues reported sensitivities ranging from 47% to 81% for Asian subgroups defined by country of origin, and 74% for Pacific Islanders in California [14]. Other studies have reported high sensitivity for the combined Asian and Pacific Islander category, including 93% reported by Clegg et al [8] and 87% by Layne et al [6].

The Utah Cancer Registry follows a workflow that initially utilizes information abstracted by hospital tumor registries from medical records as the primary source of race and ethnicity data. If the abstract received from a hospital tumor registry does not contain this information, we employ additional sources including information from other clinicians, NHIA, NAPIIA, and linkage to death and birth certificates and IHS records to fill in missing race and ethnicity information. The IHS linkage process is limited as it relies on tribal membership status which is different from racial identity.

Because central cancer registries rely heavily on hospital tumor registries for much of their race and ethnicity data, the apparent lack of standardization for healthcare systems when collecting race and ethnicity information becomes a challenge for cancer registries. Common processes used by healthcare systems for determining race and ethnicity for a cancer patient include copying the data from other available medical documentation, inferring from the patient's last name, completion at the provider's discretion, or completion by the patient or relative on an intake form [6, 20, 34, 35]. To our knowledge, the most recent detailed research regarding the source of registry race and ethnicity data in hospitals is a study done in 2003 on hospitals in the greater San Francisco Bay Area. This study assessed the proportion of hospitals in 1994 that always collect data on race and ethnicity, finding 85% and 55% respectively [36]. Hospitals reported using multiple methods to obtain race data, the most common being patient self-report (84%), patient's family (77%), patient's friend (60%), and observing patient's physical appearance (52%). A subsequent study by the same author updated statistics on the percentage of facilities reporting race and ethnicity, and proportion of hospitals using standardized forms to do so [34]. While the inclusion of patient demographic data in electronic health records as a Meaningful Use Objective may promote better data collection for race and ethnicity information, the Meaningful Use requirement is only 50% completeness [37], leaving room for improvement.

There are many valid critiques of the measurement of race and ethnicity in healthcare and research [38-42]. We recognize that as social constructs with fluid definitions, measurements of race and ethnicity are complex [43]. However, because race and Hispanic ethnicity

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

are social determinants of health that can impact patients' experiences in receiving cancer care and are correlated with cancer outcomes, collection of high-quality race and ethnicity data will remain important for understanding cancer burden that disproportionately impacts communities according to these factors [3, 41, 44].

The White House Office of Management and Budget recently proposed a revision to federal policy regarding standards for the collection of race and ethnicity data which would merge race and Hispanic ethnicity into a single question/variable, and also adds a Middle Eastern and North African race category that did not exist previously [45]. These changes, should they go into effect, will impact how cancer statistics are calculated and reported in the future.

A limitation of this study is the small percentage of individuals from races other than White within the study sample. This is not inconsistent with the population data for Utah, which as of 2020 was comprised of 79% non-Hispanic White residents [21]. Our sample was further skewed toward overrepresentation of White participants as one of the surveys included in analysis was restricted to only three cancer sites, one of which was melanoma [30], which is more prevalent in White populations [1, 3, 22]. For the AIAN race group, a population of particular interest, our study included 43 participants self-reporting as AIAN. This is a small sample but is a larger AIAN sample than any prior single-state study.

This study evaluated the validity of Utah Cancer Registry's race and ethnicity data using a large sample of participants. Findings showed that race identification is accurate for White individuals, but substantially lower for individuals of other races. These discrepancies could result in undercounting individuals of AIAN, Asian, Black, and Pacific Islander race groups. Further research is needed to understand how race and ethnicity data are collected in healthcare facilities, how variation in data collection practices affect the quality of information provided to cancer registries, and how these data can be improved.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement:

This project was supported by the U.S. Centers for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP007131. The Utah Cancer Registry is also supported by the U.S. National Cancer Institute's Surveillance, Epidemiology, and End Results Program, Contract No. HHSN261201800016I, with additional support from the University of Utah and Huntsman Cancer Foundation. This project was also supported in part by funding from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR002538. We thank Kate Hak, Lori Burke, Seth Otto, Allie Palmer, Julia Schaefer, Jenna Sawatzki, and Carlo Cardozo for their efforts to recruit participants and collect data for the studies included in this analysis.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to privacy restrictions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

References

1. Cronin KA, Scott S, Firth AU, Sung H, Henley SJ, Sherman RL, et al. (2022) Annual report to the nation on the status of cancer, part 1: National cancer statistics. Cancer 128 24:4251–84. 10.1002/cncr.34479 [PubMed: 36301149]
2. NIH National Cancer Institute (2022) Cancer. https://www.cancer.gov/about-cancer/understanding/disparities. Accessed 30 May 2023
3. Siegel RL, Miller KD, Fuchs HE, Jemal A (2022) Cancer statistics, 2022. CA: A Cancer Journal for Clinicians 72 1:7–33. 10.3322/caac.21708 [PubMed: 35020204]
4. Thornton M (2022) Standards for Cancer Registries Volume II, Data Standards and Data Dictionary. North American Association of Central Cancer Registries
5. Lin SS, Kelsey JL (2000) Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. Epidemiolic Reviews 22 2:187–202. 10.1093/oxfordjournals.epirev.a018032
6. Layne TM, Ferrucci LM, Jones BA, Smith T, Gonsalves L, Cartmel B (2019) Concordance of cancer registry and self-reported race, ethnicity, and cancer type: a report from the American Cancer Society's studies of cancer survivors. Cancer Causes & Control 30 1:21–9. 10.1007/s10552-018-1091-3 [PubMed: 30392148]
7. Altekruse SF, Cosgrove C, Cronin K, Yu M (2017) Comparing Cancer Registry Abstracted and Self-Reported Data on Race and Ethnicity. Journal of Registry Management 44 1:30–3 [PubMed: 29595942]
8. Clegg LX, Reichman ME, Hankey BF, Miller BA, Lin YD, Johnson NJ, et al. (2007) Quality of race, Hispanic ethnicity, and immigrant status in population-based cancer registry data: implications for health disparity studies. Cancer Causes & Control 18 2:177–87. 10.1007/s10552-006-0089-4 [PubMed: 17219013]
9. Jensen E, Jones N, Rabe M, Pratt B, Medina L, Orozco K, et al. (2021) The Chance That Two People Chosen at Random Are of Different Race or Ethnicity Groups Has Increased Since 2010. United States Census Bureau. https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html. Accessed 30 May 2023
10. Jones N, Bullock J. (2013) Understanding Who Reported Multiple Races in the U.S. Decennial Census: Results From Census 2000 and the 2010 Census. Family Relations 62 1:5–16. https://www.jstor.org/stable/23326022
11. Gomez SL, Glaser SL, Horn-Ross PL, Cheng I, Quach T, Clarke CA, et al. (2014) Cancer research in Asian American, Native Hawaiian, and Pacific Islander populations: accelerating cancer knowledge by acknowledging and leveraging heterogeneity. Cancer Epidemiology, Biomarkers & Prevention 23 11:2202–5. 10.1158/1055-9965.Epi-14-0624
12. Miller BA, Chu KC, Hankey BF, Ries LA (2008) Cancer incidence and mortality patterns among specific Asian and Pacific Islander populations in the U.S. Cancer Causes & Control 19 3:227–56. 10.1007/s10552-007-9088-3 [PubMed: 18066673]
13. Liu L, Tanjasiri SP, Cockburn M (2011) Challenges in identifying Native Hawaiians and Pacific Islanders in population-based cancer registries in the U.S. Journal of Immigrant and Minority Health 13 5:860–6. 10.1007/s10903-010-9381-1 [PubMed: 20803254]
14. Gomez SL, Glaser SL (2006) Misclassification of race/ethnicity in a population-based cancer registry (United States). Cancer Causes & Control 17 6:771–81. 10.1007/s10552-006-0013-y [PubMed: 16783605]
15. 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. American Journal of Epidemiology 149 11:1063–71. 10.1093/oxfordjournals.aje.a009752 [PubMed: 10355383]
16. Sweeney C, Edwards SL, Baumgartner KB, Herrick JS, Palmer LE, Murtaugh MA, et al. (2007) Recruiting Hispanic women for a population-based study: validity of surname search and characteristics of nonparticipants. American Journal of Epidemiology 166 10:1210–9. 10.1093/aje/kwm192 [PubMed: 17827445]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

17. Hamilton AS, Hofer TP, Hawley ST, Morrell D, Leventhal M, Deapen D, et al. (2009) Latinas and breast cancer outcomes: population-based sampling, ethnic identity, and acculturation assessment. Cancer Epidemiology, Biomarkers & Prevention 18 7:2022–9. 10.1158/1055-9965.Epi-09-0238

18. Silva A, Rauscher GH, Ferrans CE, Hoskins K, Rao R (2014) Assessing the quality of race/ethnicity, tumor, and breast cancer treatment information in a non-SEER state registry. Journal of Registry Management 41 1:24–30 [PubMed: 24893185]

19. Clarke LC, Rull RP, Ayanian JZ, Boer R, Deapen D, West DW, et al. (2016) Validity of Race, Ethnicity, and National Origin in Population-based Cancer Registries and Rapid Case Ascertainment Enhanced With a Spanish Surname List. Medical Care 54 1:e1–8. 10.1097/MLR.0b013e3182a30350 [PubMed: 23938598]

20. Zingmond DS, Parikh P, Louie R, Lichtensztajn DY, Ponce N, Hasnain-Wynia R, et al. (2015) Improving Hospital Reporting of Patient Race and Ethnicity--Approaches to Data Auditing. Health Services Research 50 Suppl 1 1372–89. 10.1111/1475-6773.12324 [PubMed: 26077950]

21. America Counts Staff (2021) Utah Was Fastest-Growing State From 2010 to 2020. United States Census Bureau. https://www.census.gov/library/stories/state-by-state/utah-population-change-between-census-decade.html. Accessed 4 August 2022

22. Millar M, Herget K, Codden R, Howlett C (2022) Cancer in Utah: Incidence and Mortality Statistics through 2019. Utah Cancer Registry, University of Utah. https://uofuhealth.utah.edu/documents/cancer-utah-2019

23. Hollingshaus M (2020) Utah State and County Annual Population Estimates by Single-Year of Age, Sex, and Race/Ethnicity: 2010-2019. Kem C. Gardner Policy Institute: The University of Utah. https://gardner.utah.edu/wp-content/uploads/PopEst-AgeSexRace-FS-Aug2020.pdf?x71849. Accessed 4 August 2022

24. Utah Cancer Registry Certifications & Data Quality. University of Utah Health. https://uofuhealth.utah.edu/utah-cancer-registry/about/certifications-data-quality. Accessed 30 May 2023

25. NAACCR Race and Ethnicity Work Group (2011) NAACCR Guildeline for Enhancing Hispanic/Latino Identification: Revised NAACCR Hispanic/Latino Identification Algorithm [NHIA v2.2.1]. North American Association of Central Cancer Registries.

26. NAACCR Race and Ethnicity Work Group (2011) NAACCR Asian/Pacific Islander Identification Algorithm [NAPIIA v1.2.1]: Enhancing the Specificity of Identification. North American Associatin of Central Cancer Registries.

27. Soisson S, Ganz PA, Gaffney D, Rowe K, Snyder J, Wan Y, et al. (2018) Long-term Cardiovascular Outcomes Among Endometrial Cancer Survivors in a Large, Population-Based Cohort Study. Journal of the National Cancer Institute 110 12:1342–51. 10.1016/j.ygyno.2017.12.025 [PubMed: 29741696]

28. Millar MM, Herget KA, Ofori-Atta B, Codden RR, Edwards SL, Carter ME, et al. (2023) Cancer survivorship experiences in Utah: an evaluation assessing indicators of survivors' quality of life, health behaviors, and access to health services. Cancer Causes & Control 34 4:337–47. 10.1007/s10552-023-01671-5 [PubMed: 36723708]

29. Millar MM, Elena JW, Gallicchio L, Edwards SL, Carter ME, Herget KA, et al. (2019) The feasibility of web surveys for obtaining patient-reported outcomes from cancer survivors: a randomized experiment comparing survey modes and brochure enclosures. BMC Medical Research Methodology 19 1:208. 10.1186/s12874-019-0859-9 [PubMed: 31730474]

30. McCarty R, Trabert B, Millar M, Haaland B, Grieshober L, Barnard M, et al. Abstract 6471: Tattooing and risk of hematologic cancer: A population-based case-control study in Utah. Cancer Research; 2023. 10.1158/1538-7445.AM2023-6471

31. U.S. Department of Agriculture Economic Research Service (2020) Rural-Urban Continuum Codes. https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/. Accessed 30 May 2023

32. Parker K, Horowitz J, Morin R, Lopez M (2015) Chapter 7: The Many Dimensions of Hispanic Racial Identity. Multiracial in America. Pew Research Center. https://www.pewresearch.org/social-trends/2015/06/11/chapter-7-the-many-dimensions-of-hispanic-racial-identity/. Accessed 30 May 2023

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

33. Kem C Gardner Policy Institute. 2020 Census Redistricting Data. https://gardner.utah.edu/demographics/2020-census/redistricting-data. Accessed 30 May 2023.

34. Gomez SL, Lichtensztajn DY, Parikh P, Hasnain-Wynia R, Ponce N, Zingmond D (2014) Hospital practices in the collection of patient race, ethnicity, and language data: a statewide survey, California, 2011. Journal of Health Care for the Poor and Underserved 25 3:1384–96. 10.1353/hpu.2014.0126 [PubMed: 25130247]

35. Gomez SL, Satariano W, Le GM, Weeks P, McClure L, West DW (2009) Variability among hospitals and staff in collection of race, ethnicity, birthplace, and socioeconomic information in the greater San Francisco Bay Area. Journal of Registry Management 36 4:105–10 [PubMed: 20795551]

36. Gomez SL, Le GM, West DW, Satariano WA, O'Connor L (2003) Hospital policy and practice regarding the collection of data on race, ethnicity, and birthplace. American Journal of Public Health 93 10:1685–8. 10.2105/ajph.93.10.1685 [PubMed: 14534222]

37. Blumenthal D, Tavenner M (2010) The "meaningful use" regulation for electronic health records. N Engl J Med 363 6:501–4. 10.1056/NEJMp1006114 [PubMed: 20647183]

38. Magaña López M, Bevans M, Wehrlen L, Yang L, Wallen GR (2016) Discrepancies in Race and Ethnicity Documentation: a Potential Barrier in Identifying Racial and Ethnic Disparities. Journal of Racial and Ethnic Health Disparities 4 5:812–8. 10.1007/s40615-016-0283-3 [PubMed: 27631381]

39. Cook LA, Sachs J, Weiskopf NG (2021) The quality of social determinants data in the electronic health record: a systematic review. Journal of the American Medical Informatics Association 29 1:187–96. 10.1093/jamia/ocab199. [PubMed: 34664641]

40. Webster PS, Fulton JP, Sampangi S (2013) Conflicting race/ethnicity reports: lessons for improvement in data quality. Journal of Registry Management 40 3:122–6 [PubMed: 24643214]

41. Webster PS, Sampangi S (2017) Did We Have an Impact? Changes in Racial and Ethnic Composition of Patient Populations Following Implementation of a Pilot Program. Journal of Healthcare Quality 39 3:e22–e32. 10.1111/jhq.12079

42. Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S (2005) Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. Annals of Epidemiology 15 1:71–9. 10.1016/j.annepidem.2004.03.002 [PubMed: 15571996]

43. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, et al. (2015) Accuracy of race, ethnicity, and language preference in an electronic health record. Journal of General Internal Medicine 30 6:719–23. 10.1007/s11606-014-3102-8 [PubMed: 25527336]

44. Flanagin A, Frey T, Christiansen SL (2021) Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. Journal of the American Medical Association 326 7:621–7. 10.1001/jama.2021.13304 [PubMed: 34402850]

45. Management and Budget Office (2023) Initial Proposals For Updating OMB's Race and Ethnicity Statistical Standards. https://www.federalregister.gov/documents/2023/01/27/2023-01635/initial-proposals-for-updating-ombs-race-and-ethnicity-statistical-standards. Accessed 30 May 2023.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1:

Demographic characteristics of total study sample and by survey study

	Total (N=3162)		Study 1 (N=443)		Study 2 (N=202)		Study 3 (N=1512)		Study 4 (N=1005)	
	n	%	n	%	n	%	n	%	n	%
Survey Year(s)			2015,2018		2017		2019-2022		2020-2022	
Sex										
Male	1539	48.7	219	49.4	81	40.1	698	46.2	541	53.8
Female	1623	51.3	224	50.6	121	59.9	814	53.8	464	46.2
Age at survey										
<45	394	12.5	13	2.9	49	24.3	129	8.5	203	20.2
45-54	408	12.9	44	9.9	64	31.7	169	11.2	131	13.0
55-64	786	24.9	132	29.8	72	35.6	359	23.7	223	22.2
65-74	1036	32.8	184	41.5	17	8.4	498	32.9	337	33.5
75+	538	17.0	70	15.8	0	0.0	357	23.6	111	11.0
Cancer site										
Breast	370	11.7	0	0.0	46	22.8	324	21.4	0	0.0
Endometrial	251	7.9	184	41.5	0	0.0	67	4.4	0	0.0
Leukemia	163	5.2	0	0.0	0	0.0	29	1.9	134	13.3
Lymphoma	412	13.0	0	0.0	0	0.0	72	4.8	340	33.8
Melanoma	733	23.2	0	0.0	0	0.0	202	13.4	531	52.8
Oral Cavity	288	9.1	259	58.5	0	0.0	29	1.9	0	0.0
Prostate	329	10.4	0	0.0	39	19.3	290	19.2	0	0.0
All other cancers	616	19.5	0	0.0	117	57.9	499	33.0	0	0.0
Diagnosis year										
2012 or earlier	384	12.1	311	70.2	73	36.1	0	0.0	0	0.0
2013-2015	653	20.7	94	21.2	46	22.8	513	33.9	0	0.0
2016-2018	962	30.4	38	8.6	83	41.1	841	55.6	0	0.0
2019-2021	1163	36.8	0	0.0	0	0.0	158	10.4	1005	100.0
Geography										
Urban	2762	87.3	383	86.5	173	85.6	1330	88	876	87.2
Rural	400	12.7	60	13.5	29	14.4	182	12.0	129	12.8

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2:

Agreement between Cancer Registry Race and Ethnicity Classifications and Gold-standard Self-report

	Self-Report: Yes		Registry Report: Yes		Sensitivity		Specificity		Positive Predictive Value	
	n	%[a]	n	%[a]	Value	(95% CI[b])	Value	(95% CI)	Value	(95% CI)
Ethnicity [c]										
Hispanic	323	10.2	383	12.1	88.2	(84.2 - 91.5)	96.5	(95.8 - 97.2)	74.4	(69.7 - 78.7)
Race [d]										
American Indian or Alaska Native	43	1.4	11	0.3	9.3	(2.6 - 22.1)	99.8	(99.5 - 99.9)	36.4	(10.9 - 69.2)
Asian	32	1.0	20	0.6	40.6	(23.7 - 59.4)	99.8	(99.5 - 99.9)	65.0	(40.8 - 84.6)
Black or African American	23	0.7	15	0.5	60.9	(38.5 - 80.3)	100.0	(99.8 - 100.0)	93.3	(68.1 - 99.8)
Pacific Islander	16	0.5	^	^	25.0	(7.3 - 52.4)	99.9	(99.8 - 100.0)	66.7	(22.3 - 95.7)
White	2994	94.7	3009	95.2	98.8	(98.3 - 99.2)	47.4	(37.2 - 57.8)	98.3	(97.8 - 98.7)

^ Small cell counts (<=10 are suppressed for confidentiality)

[a] Percent of participants included in analysis, n=3,162.

[b] CI, Confidence Interval

[c] Individuals who did not provide a response to the Hispanic ethnicity survey question (n=23) are excluded from this table and from the calculations for Hispanic ethnicity sensitivity and specificity.

[d] Individuals who did not provide any self-reported race (missing race n=71) are excluded from this table and for calculations for sensitivity and specificity for each race group. Individuals who reported their race as “Other” (n=29) are included in sensitivity and specificity calculations, but we did not calculate sensitivity and specificity for Other race.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

Agreement between Cancer Registry American Indian or Alaska Native Classification and Self-reported AIAN classification by Hispanic ethnicity, age, and diagnosis year

	Self-Report of AIAN: Yes[a]	Sensitivity		Specificity		Positive Predictive Value	
		Value	**(95% CI[b])**	**Value**	**(95% CI)**	**Value**	**(95% CI)**
Ethnicity							
Self-Reported Hispanic	23	4.3	(0.1 - 21.9)	99.1	(97.0 - 99.9)	33.3	(0.8 - 90.6)
Self-Reported Non-Hispanic	19	10.5	(1.3 - 33.1)	99.8	(99.6 - 99.9)	28.6	(3.7 - 71.0)
Age at Survey							
<45	^	12.5	(0.3 - 52.7)	99.7	(98.5 - 100.0)	50.0	(1.3 - 98.7)
45-54	^	0.0	(0.0 - 45.9)	100.0	(99.1 - 100.0)	-	-
55-64	^	11.1	(0.3 - 48.2)	99.5	(98.6 - 99.9)	20.0	(0.5 - 71.6)
65-74	13	7.7	(0.2 - 36.0)	99.9	(99.4 - 100.0)	50.0	(1.3 - 98.7)
75+	^	14.3	(0.4 - 57.9)	99.8	(98.9 - 100.0)	50.0	(1.3 - 98.7)
Diagnosis year							
2012 or earlier	^	33.3	(0.8 - 90.6)	99.7	(98.5 - 100.0)	50.0	(1.3 - 98.7)
2013-2015	^	0.0	(0.0 - 36.9)	99.8	(99.1 - 100.0)	0.0	(0.0 - 97.5)
2016-2018	21	14.3	(3.0 - 36.3)	99.6	(98.9 - 99.9)	42.9	(9.9 - 81.6)
2019-2021	11	0.0	(0.0 - 28.5)	99.9	(99.5 - 100.0)	0.0	(0.0 - 97.5)

^ Small cell counts (<=10 are suppressed for confidentiality)

[a] Due to small cell sizes, we are unable to also display counts for Registry Report: Yes in this table.

[b] CI, Confidence Interval