

# How generalizable are the SEER registries to the cancer populations of the USA?

Tzy-Mey Kuo<sup>1</sup>  · Lee R. Mobley<sup>2</sup>

Received: 26 February 2016 / Accepted: 16 July 2016 / Published online: 21 July 2016  
© Springer International Publishing Switzerland 2016

## Abstract

**Background** We determined whether the current SEER registries are representative of the nation's cancer cases or the socio-demographic characteristics.

**Methods** We used breast cancer (BC) and colorectal cancer (CRC) cases diagnosed 2004–2009 from the US Cancer Statistics (USCS) database. Cases were classified into groups residing in SEER coverage areas and the other areas. We compared difference between SEER and non-SEER areas in: age–race-specific proportions of late-stage BC or CRC, area demographics and socioeconomic factors, and data quality.

**Results** For late-stage CRC diagnosis, SEER areas contained lower proportions of people with other race and higher proportions of Asian and Hispanic females aged <40, than non-SEER areas. For late-stage BC diagnosis, SEER and non-SEER estimates were comparable. SEER areas had lower percentages of whites and higher percentages of young people, were more urban, and had higher percentage of poor, lower educational attainment, and higher unemployment. SEER areas also tended to have a higher percentage of case completeness than non-SEER areas.

**Conclusion** Overall, SEER registries were not significantly different from non-SEER areas in terms of average age–race-specific proportions of late-stage BC or CRC, except

for estimates of late-stage CRC for other race and young Asian and Hispanic women. Although case completeness was better in SEER areas than non-SEER areas, SEER areas had greater economic disadvantage and greater minority diversity among the population. This study demonstrated a need for caution in using SEER data and discussed advantages of using the more complete USCS database.

**Keywords** SEER · Generalizable · Registry · Cancer · Late stage

## Introduction

The Surveillance, Epidemiology, and End Results (SEER) program, originally funded by the National Cancer Institute (NCI) in 1973, collects information of cancer incidence and survival from strategically selected cancer registries of the USA. The SEER program began by including nine registries and has expanded over time to now include 18 registries. Currently, SEER registries cover 28 % of the US population [1]. SEER registries have been the foundation for decades of population-based cancer-related research. In addition, SEER–Medicare linked databases are also widely used for cancer studies related to the population age 65 and older [2]. These data have been available for decades to researchers under certain data use agreements and have been used in numerous peer-reviewed studies.

By contrast, fewer researchers have used the more recently available United States Cancer Statistics (USCS) database of cancer registries that combines data from the National Program of Cancer Registries (NPCR) and SEER registries. The NPCR was funded by CDC since 1992 to support central cancer registries and covers 96 % of the US

✉ Tzy-Mey Kuo  
tkuo@email.unc.edu

<sup>1</sup> Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, CB 7293, Chapel Hill, NC 27599-7293, USA

<sup>2</sup> School of Public Health and Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA, USA

population including 45 states, the District of Columbia, Puerto Rico, and the US Pacific Island Jurisdictions [3]. The combination of databases from SEER and NPCR collects data from the entire nation that provides a single, pooled-state database of reconciled, comparable cancer information geocoded at the local level to facilitate cancer control planning and evaluation [4]. Using these data, CDC and NCI, in collaboration with the North American Association of Central Cancer Registries (NAACCR), certify registries that meet high data quality criteria and jointly publish official federal statistics on cancer incidence from registries that meet the USCS publication standard [5]. This comprehensive database is now available inside National Centers for Health Statistics (NCHS) and Census Research Data Centers (RDCs) to qualified researchers [6]. This combined database has been used by only a few researchers because it has not been broadly available until recently [7].

With the newly available USCS database, it is important to know if studies using this database produce similar results to the estimates from the widely used SEER data, which is a subset of it. In preparing the joint publication from NCI and CDC for the official federal cancer statistics, Wingo et al. [4] used USCS data in 1999 to compare the cancer incidence rates from SEER and NPCR data. They found that cancer incidence rates for specific sites varied by registries and that the SEER under-represented rates for all sites combined. They also found that the SEER over-represented the rate for breast cancer (BC) but under-represented the rate for colorectal cancer (CRC).

In addition, Henley et al. [7] used the USCS data before they became publicly available to examine cancer incidence during 2004–2006, for persons aged >50 at time of diagnosis. They used the SEER summary stage [8] to define late stage as including both regional and distant stages. The study examined two outcomes: age-adjustment incidence rate per 100,000 of the general population and proportions of cancer cases that were diagnosed at late stage. Henley et al. [7] found that age-adjusted incidence rates of late-stage cancer differed by age groups and by race or ethnicity. Rates of late-stage CRC incidence increased with age and were highest among black men and women. Late-stage BC incidence rates increased with age only through age 79 and were highest among the age 60–79 group and black women. The study also showed that roughly a third of new BC cases and roughly half of CRC cases were diagnosed at a late stage in the USA between the years 2004–2006. Similarly, Mobley et al. [9] using SEER registry cancer population data from 11 states (CA, CT, GA, IA, KY, LA, MI, NJ, NM, WA, and UT) over 2000–2005 found comparable results for the proportions of late-stage BC and CRC. Specifically, Mobley et al. found that proportion of late-stage CRC ranged from 49 to 58 %, while

BC ranged from 26 to 33 % across the 11 SEER registry states.

In addition, other studies compared cancer mortality rates between US and SEER registries [10, 11]. For example, Frey et al. [10] used the age-adjusted mortality rate at the county level in 1975–1988 and compared race-, sex-, and cancer site-specific trends and levels between the SEER areas and the entire nation. Overall, although most of the trends and levels of mortality rates were similar between the two areas, some comparisons showed that SEER coverage areas were not representative of the US population. Specifically, SEER areas over-represented BC mortality rates in black female populations and under-represented CRC mortality rates for both white male and female populations. Using more recent data (i.e., 1992–2000), Merrill and Dearden [11] also found that SEER registries under-represented US cancer mortality for CRC in white males and females and black males. In general, they found SEER tended to under-represent the US mortality in several cancer sites and across the majority of the SEER registries, with the worst under-representation in Utah and New Mexico.

The geographic areas in the SEER registries were strategically selected by the SEER program to ensure a high-quality cancer reporting system and to include epidemiologically significant population subgroups [1]. What is not known is the extent to which SEER regions are representative of the nation in terms of population demography and geographic socioeconomic characteristics. This is the main contribution of this paper, which is important because such factors have been demonstrated to relate strongly and significantly to cancer incidence and stage of diagnosis [9, 12–15].

Two rather dated studies have addressed the socioeconomic and demographic representativeness of the SEER population. One focused on the general population [16] while the other focused on the elderly population [2]. Nattinger et al. [16] used 1990 data to examine the generalizability of SEER for factors relevant to epidemiologic and healthcare research by classifying 198 counties as the SEER coverage areas and 2882 as the non-SEER areas. They found the SEER areas to be more urban, more affluent, and with lower unemployment rates than the USA as a whole.

The Warren et al. [2] study compared socio-demographic characteristics between persons 65 years and older in the SEER coverage counties and the US population age 65 and older, using 1990 census data. They found that the age and sex distributions for individuals 65 years and older in the SEER areas were comparable with that of the US elderly population. However, the SEER areas had a lower proportion of white persons and a higher proportion of persons of other races (persons who were not white,

African-American, or unknown race). In addition, these older persons in the SEER areas were much more likely to have resided in an urban setting, a more affluent area, or an area with higher managed care penetration among the Medicare enrollee population than the average US residents age 65 years and older.

Our paper seeks to make an important contribution, because the two studies regarding the representativeness of the SEER registry populations and their geographic settings are quite dated [2, 16]. Many socioeconomic and demographic changes have occurred in the USA over the past 20 years. According to the US Census Bureau the percentage of white population decreased from 80.3 % in 1990 to 72.4 % in 2010, whereas the percentage of black population only increased slightly (12.1 to 12.6 %) and the percentage increased from 7.6 to 12.1 % for the other nonwhite race populations [17]. The entire population is also aging, with an increasing percentage of older people and a decreasing percentage of younger people as time passes [18]. Also, there was a decline in the middle class with the Great Recession during 2007–2009, which affected the mix of socioeconomics across the nation. In addition, the SEER registries have expanded to include all of Georgia and California as compared to including only the Atlanta and San Francisco areas in these states, which was the SEER coverage for the study by Nattinger et al. [16]. Furthermore, the states of Kentucky, Louisiana, and New Jersey also joined the SEER program after 2000. Thus, the SEER areas today include many more rural areas and minority populations than the earlier SEER areas. What is not known is whether the cancer population data obtained from SEER registries today are still generalizable and representative of the entire US cancer population, and whether the geographic settings of the SEER areas are representative of the USA, in terms of population demographics and socio-ecological factors. Finally, it is well documented that the SEER registries were originally selected for better data quality as compared to the other cancer registries. Over the years CDC, NCI, and NAACCR have worked together to help all cancer registries achieve standards that reflect high-quality data. It is unknown to what extent the data quality differs between SEER and non-SEER registries in current data.

The purpose of this study is to use the newly available USCS database and stratify the data into SEER areas and non-SEER areas, to determine to what extent the SEER registries are generalizable to the nation, and in which dimensions they are not fully generalizable. Specifically, this paper provides an update and current comparison of the cancer cases, demographics, socioeconomic factors, and data quality in SEER areas as compared with the rest of the USA. For the cancer case comparison, we examined the comparability of late-stage cancer diagnosis, focusing on

BC and CRC, the two cancers that are preventable through screening but remain among the top four most commonly diagnosed cancers in the USA [19]. We also compared population demographics and socioeconomic factors in SEER and non-SEER areas. For dimensions of data quality, we focus on the key components of (1) completeness of case ascertainment and (2) missing data on race, sex, and age. This study answers the following questions: Are the SEER registries representative of the US cancer population in terms of late-stage BC and CRC diagnoses? Are they representative of the current socioeconomic and demographic characteristics which underpin the observed health outcomes in national statistics? Is the data quality comparable between SEER registries and the rest of the registries in the USA?

## Methods

### Study sample

We obtained data from the USCS Restricted Access Dataset and selected people of all ages with a first breast or colorectal cancer diagnosed in 2004–2009. BC was identified by the third edition of International Classification of Diseases for Oncology (ICD-O-3) code of C50, and CRC was identified by the ICD-O-3 codes of C18, C19.9, C20.9, and C26.0 [20]. The dataset was prepared by NCHS, whereby BC or CRC that is not the primary cancers, records with unknown cancer stage or unstaged cancer in the SEER summary stage variable, and missing data in age, race, or gender were excluded. From this dataset, we further excluded records when diagnosis was by autopsy or from death certificates (<1 % of all cases). For BC, we also excluded males. Using the SEER summary stage 2000 variable provided in the USCS database, we coded regional or distant diagnosis as late stage, and in situ or localized diagnosis as early stage. We then created a binary indicator for each individual specifying whether their cancer was diagnosed at a late stage or not.

### Definition of SEER and non-SEER areas

Three states (Kansas, Maryland, and Minnesota) were not present in the USCS database over the timeframe of this study. We excluded an additional state, Virginia, because data were not available until 2007. Among the included 46 states, 10 are completely covered by SEER registries (HI, CA, UT, NM, IA, KY, GA, CT, NJ, and LA) and two others (WA and MI) have partial coverage over some counties by SEER registries (including Clallam, Grays Harbor, Island, Jefferson, King, Kitsap, Mason, Pierce, San Juan, Skagit, Snohomish, Thurston, and Whatcom counties

in WA; Macomb, Oakland, and Wayne counties in MI). To provide a clear comparison for the population characteristics between SEER and non-SEER areas, we focused on the SEER areas that represent entire populations; thus, the SEER registries in Alaska, Oklahoma, and Arizona that cover native Americans and Alaska Natives only were not counted as ‘SEER areas’ in our study. Therefore, we define SEER areas as the above-mentioned 10 SEER states and the counties of WA and MI that are covered by SEER registries. The rest of the nation’s counties were defined as non-SEER areas. Figure 1 shows the geographic areas of the SEER and non-SEER areas in this study. We assigned cancer patients to the appropriate group based on their county or state of residence at the time of diagnosis.

To assess the representativeness of SEER registries for the nation, we first compared the similarity of cancer burden between SEER and non-SEER areas. Specifically, we compared the proportions of late-stage BC and CRC diagnosis between SEER and non-SEER coverage areas, across subgroups of age and racial or ethnic groups. We defined five age groups (age <40, 40–49, 50–64, 65–74, and 75+) and five racial or ethnic groups (white, African-American, Asian, Hispanic, and other). We also provided comparisons for each racial or ethnic group by the five age groups.

Next, we assessed the comparability of the underlying population demographics and socioeconomic characteristics measured at county level, using census 2000 and 2010 data. We identified 612 counties in the SEER areas and 2529 (or 2531 in year 2010) counties in the non-SEER areas. We then compared county-level population demographic information, poverty, health insurance, educational attainment, unemployment rate, and percent of population living in urban areas, between SEER areas and non-SEER areas. Demographic data in 2000 were obtained from the Area Health Resource Files (AHRF), whereas the 2010 data were obtained from census 2010; poverty data were from the Small Area Income and Poverty Estimates (SAIPE); health insurance information was from the Small Area Health Insurance Estimates (SAHIE). Data for educational attainment and the unemployment rate in 2000 were available in the AHRF, whereas these variables were not available from AHRF for 2010. For the comparison of these two variables, we used the similar estimates from the 2005–2009 American Community Survey (ACS) five-year estimates at county level.

For the data quality comparison, we obtained Data Quality Indicators published from Cancer in North America (CINA) for cancer data from 2008 to 2012 [21] to compare case completeness and missing data on race, sex, and age variables. Registries that did not meet publication criteria as indicated in the CINA data were excluded from our analysis (Minnesota in 2008–2012, Arkansas in 2010, and

Nevada in 2011–2012 were excluded from our analyses). Because the data quality criteria were evaluated annually, we compared the data quality between SEER and non-SEER areas separately for each year from 2008 to 2012 (the most recent data). NAACCR defines the case completeness indicator for each registry as the percentage derived from dividing observed age-adjusted cancer incidence rate by the expected age-adjusted incidence rate, which is a weighted summation of race-specific case completeness ascertainment for white and black races [22]. Missing data on race, sex, and age were defined as the percentage of the data fields that have missing or unknown values.

### Statistical analyses

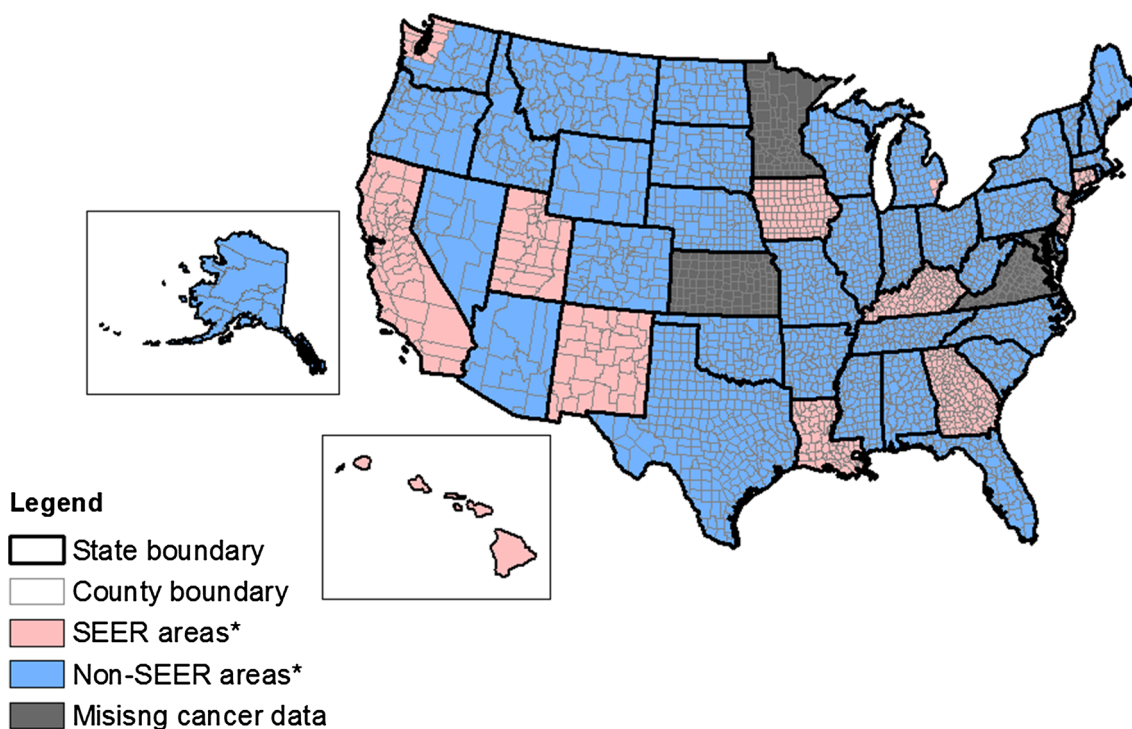
We performed three sets of comparisons, and the analytic units are different for the comparisons. The analysis concerning the difference in late-stage BC and CRC between SEER and non-SEER areas was based on person-level data; the comparison of area-level socio-demographics was based on county-level data; the comparison of registry data quality was based on registry-level data. For the analyses concerning late-stage BC and CRC and area-level socio-demographics, the sample sizes are large and reach statistical significance in conventional tests even with very small differences. Instead of using statistical significance tests to evaluate the differences between SEER and non-SEER areas, we used the ‘standardized difference’ to quantify the similarity (or differences). The standardized difference approach takes the difference in the variables of interest between the SEER and non-SEER areas and divides it by the pooled standard deviation of the two groups (see the equations below). Thus, the standardized difference estimate is expressed per unit of standard deviation, which is scale-free and less sensitive to sample size. The standardized difference is often called the ‘effect size’ in social science and has been widely used in comparative effectiveness research of medical studies to compare the balance of treatment and control groups in order to assess treatment effect using observational data [23].

In this study, we applied the following formula to compute standardized difference for dichotomous variables (whether a person had late-stage cancer diagnosis or not) in Eq. (1) and continuous variables (county-level characteristics and data quality) in Eq. (2).

Standardized difference between two proportions

$$= \frac{(P1 - P2)}{\sqrt{(P1(1 - P1) + P2(1 - P2))/2}} \quad (1)$$

where  $P1$  is the average proportion of late-stage cancer diagnosis for people in SEER areas and  $P2$  is the same measure for people in non-SEER areas.



\* In the study we defined SEER and non-SEER areas as states/counties where registries represent the entire populations (see methods section for details).

**Fig. 1** SEER areas and non-SEER areas in the USA

Standardized difference between two continuous variables

$$= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2 + s_2^2)/2}} \tag{2}$$

where  $\bar{x}_i$  is the average score and  $s_i^2$  is the variance, for SEER areas and non-SEER areas indicated as i.

Because the samples of SEER areas and non-SEER areas differ in size, the pooled standard deviation (i.e., denominator of Eq. 2) can be estimated using

$\sqrt{\frac{(n1-1)s_1^2 + (n2-1)s_2^2}{n1+n2-2}}$  where ni is the sample size for SEER and non-SEER groups.

Although the standardized difference provides an intuitive estimate for comparability between groups, there is no consensus about what threshold of standardized difference indicates imbalance or meaningful difference. An absolute value of the standardized difference of <0.1 (an effect that is <1/10 of a standard deviation in the covariate of interest) is often suggested by researchers to indicate a negligible difference in proportions or means between groups [23]. Therefore, we used the 0.1 criterion for the standardized difference as a cutoff point to evaluate our findings. All analyses were conducted using SAS software, version 9.3 (SAS Institute, Cary, NC, USA).

**Results**

We identified 1,151,096 cases of BC and 658,470 cases of CRC in the 46 states with data available in 2004–2009. The SEER registry coverage areas contained less than one-third of BC or CRC total cancer cases. Among the total cancer cases, 30.8 % of BC cases and 54.2 % of CRC cases were diagnosed at late stage. The proportions of late-stage diagnosis differed slightly between SEER and non-SEER areas for both cancers: 30.7 % in SEER areas and 30.9 % in non-SEER areas for BC; 55.2 % and 53.8 % for CRC, respectively.

The comparisons of the proportion of late-stage BC or CRC by age groups and by race or ethnicity are presented in Table 1. Among all the 105 comparisons, 10 have absolute value of standardized difference greater than 0.1, which are considered not comparable. Thus, in the vast majority of the comparisons, we found that the proportions of late-stage BC or CRC diagnoses were comparable between the cancer populations in the SEER and non-SEER coverage areas. The 10 non-comparable groups are only found in people with CRC. Specifically, for both males and females with CRC, the SEER coverage areas had a meaningfully lower proportion of late-stage diagnoses than the non-SEER areas for people in ‘other’ races.

**Table 1** Comparisons of proportions of late-stage CRC and BC diagnoses across SEER and non-SEER coverage areas, by race or ethnicity and age

	Males—CRC			Females—CRC			Females—BC		
	SEER	Non-SEER	Std. diff. <sup>a</sup>	SEER	Non-SEER	Std. diff. <sup>a</sup>	SEER	Non-SEER	Std. diff. <sup>a</sup>
Age group									
<40	0.66	0.65	0.021	0.64	0.62	0.041	0.46	0.46	0
40–49	0.62	0.63	−0.021	0.62	0.61	0.021	0.33	0.33	0
50–64	0.53	0.53	0	0.55	0.54	0.02	0.30	0.31	−0.022
65–74	0.53	0.51	0.04	0.54	0.52	0.04	0.27	0.27	0
75 plus	0.54	0.52	0.04	0.56	0.53	0.06	0.29	0.29	0
Race/ethnicity									
White	0.54	0.53	0.02	0.56	0.54	0.04	0.29	0.30	−0.022
African-American	0.57	0.57	0	0.56	0.55	0.02	0.38	0.39	−0.021
Asian	0.56	0.57	−0.02	0.58	0.58	0	0.28	0.30	−0.044
Hispanic	0.57	0.57	0	0.58	0.58	0	0.36	0.36	0
All others	<b>0.32</b>	<b>0.39</b>	<b>−0.147</b>	<b>0.34</b>	<b>0.42</b>	<b>−0.165</b>	0.28	0.28	0
White, by age									
<40	0.64	0.64	0	0.60	0.61	−0.02	0.44	0.44	0
40–49	0.62	0.63	−0.021	0.62	0.61	0.021	0.31	0.32	−0.022
50–64	0.53	0.53	0	0.55	0.55	0	0.29	0.29	0
65–74	0.52	0.51	0.02	0.54	0.52	0.04	0.26	0.26	0
75 plus	0.54	0.51	0.06	0.56	0.53	0.06	0.28	0.28	0
African-American, by age									
<40	0.68	0.69	−0.022	0.63	0.64	−0.021	0.50	0.52	−0.04
40–49	0.65	0.66	−0.021	0.62	0.63	−0.021	0.40	0.42	−0.041
50–64	0.56	0.56	0	0.55	0.54	0.02	0.37	0.38	−0.021
65–74	0.56	0.54	0.04	0.53	0.53	0	0.33	0.34	−0.021
75 plus	0.56	0.55	0.02	0.58	0.55	0.061	0.37	0.36	0.021
Asian, by age									
<40	0.70	0.68	0.043	<b>0.76</b>	<b>0.64</b>	<b>0.264</b>	0.40	0.40	0
40–49	0.65	0.65	0	0.64	0.67	−0.063	0.30	0.28	0.044
50–64	0.55	0.55	0	0.55	0.55	0	0.28	0.30	−0.044
65–74	0.52	0.56	−0.08	0.56	0.56	0	0.24	0.27	−0.069
75 plus	0.57	0.55	0.04	0.58	0.58	0	0.27	0.28	−0.022
Hispanic, by age									
<40	0.72	0.69	0.066	<b>0.70</b>	<b>0.65</b>	<b>0.107</b>	0.52	0.49	0.06
40–49	0.63	0.64	−0.021	0.63	0.64	−0.021	0.38	0.37	0.021
50–64	0.55	0.57	−0.04	0.56	0.56	0	0.34	0.35	−0.021
65–74	0.56	0.55	0.02	0.55	0.55	0	0.31	0.30	0.022
75 plus	0.55	0.54	0.02	0.57	0.57	0	0.31	0.32	−0.022
All others, by age									
<40	<b>0.35</b>	<b>0.52</b>	<b>−0.348</b>	<b>0.45</b>	<b>0.53</b>	<b>−0.161</b>	0.42	0.40	0.041
40–49	0.40	0.42	−0.041	<b>0.35</b>	<b>0.44</b>	<b>−0.185</b>	0.29	0.31	−0.044
50–64	<b>0.27</b>	<b>0.39</b>	<b>−0.257</b>	<b>0.24</b>	<b>0.37</b>	<b>−0.285</b>	0.26	0.28	−0.045
65–74	<b>0.31</b>	<b>0.37</b>	<b>−0.127</b>	0.39	0.43	−0.081	0.26	0.24	0.046
75 plus	0.41	0.39	0.041	0.44	0.44	0	0.26	0.23	0.070

Numbers highlighted in bold are considered non-negligible difference

<sup>a</sup> Standardized difference

For this racial group, the pattern was found in all age groups except for age 40–49, 75 years and older males, and 65 years and older females. For females with CRC, we also found Asian and Hispanic women aged <40 had a meaningfully higher proportion of late-stage diagnosis in SEER areas than in non-SEER areas. For all women with BC, our data suggested that the proportion of late-stage diagnosis was comparable between SEER and non-SEER areas.

Table 2 presents the results of the analysis of whether the demographic and socioeconomic characteristics were similar between the SEER and non-SEER areas in both 2000 and 2010. As shown in Table 2, overall we found the patterns (i.e., standardized difference between SEER and non-SEER areas) were similar and consistent in both 2000

and 2010. Using the 0.1 criterion as the cutoff for a meaningful difference, we found SEER areas had higher percentages of younger (<age 44) population than the rest of US areas did, in both years. There were also smaller percentage of whites and American Indians and a higher percentage of African-Americans and Asians in the general population in SEER areas, as compared to the non-SEER areas. SEER areas also had a higher percentage of the population in poverty, lower educational attainment, and higher unemployment rate, compared to non-SEER areas. For the comparison of health insurance, the SEER areas had a higher percentage of uninsured in 2000, compared to the non-SEER areas. However, the percentage of uninsured population increased and became equivalent in both areas

**Table 2** Comparisons in county-level socio-demographic characteristics between SEER and non-SEER areas, in 2000 and 2010

Variables	2000			2010		
	SEER ( <i>N</i> = 612)	Non-SEER ( <i>N</i> = 2529)	Std. diff. <sup>c</sup>	SEER ( <i>N</i> = 612)	Non-SEER ( <i>N</i> = 2531)	Std. diff. <sup>c</sup>
Gender <sup>a</sup>						
Male	49.5	49.6	−0.054	49.8	50.0	−0.091
Age groups <sup>a</sup>						
Age 44 or younger	<b>63.6</b>	<b>61.4</b>	<b>0.372</b>	<b>57.8</b>	<b>55.4</b>	<b>0.359</b>
Age 45–54	<b>13.5</b>	<b>13.7</b>	<b>−0.128</b>	<b>14.7</b>	<b>15.0</b>	<b>−0.187</b>
Age 55–64	<b>9.4</b>	<b>9.8</b>	<b>−0.277</b>	<b>12.8</b>	<b>13.3</b>	<b>−0.243</b>
Age 65–74	<b>7.1</b>	<b>7.8</b>	<b>−0.354</b>	<b>8.1</b>	<b>8.8</b>	<b>−0.317</b>
Age 75 plus	<b>6.4</b>	<b>7.2</b>	<b>−0.346</b>	<b>6.6</b>	<b>7.4</b>	<b>−0.368</b>
Race and ethnicity <sup>a</sup>						
White	<b>80.2</b>	<b>85.4</b>	<b>−0.314</b>	<b>78.4</b>	<b>83.9</b>	<b>−0.329</b>
African-American	<b>12.3</b>	<b>7.9</b>	<b>0.304</b>	<b>12.5</b>	<b>8.0</b>	<b>0.314</b>
American Indians	<b>1.1</b>	<b>2.1</b>	<b>−0.138</b>	<b>1.1</b>	<b>2.2</b>	<b>−0.148</b>
Asian	<b>1.5</b>	<b>0.7</b>	<b>0.369</b>	<b>1.9</b>	<b>1.0</b>	<b>0.362</b>
Hispanic	7.0	6.0	0.084	9.3	8.0	0.093
Poverty <sup>b</sup>						
People living in poverty	<b>14.3</b>	<b>13.0</b>	<b>0.254</b>	<b>18.2</b>	<b>16.4</b>	<b>0.284</b>
Insurance <sup>c</sup>						
Uninsured population	<b>15.6</b>	<b>14.6</b>	<b>0.208</b>	18.7	18.5	0.033
Medicare managed care penetration rate	<b>6.0</b>	<b>4.7</b>	<b>0.141</b>	14.8	15.4	−0.052
Educational attainment <sup>d</sup>						
Population aged 25+ years with less than a high school diploma or equivalent	<b>10.3</b>	<b>8.8</b>	<b>0.295</b>	<b>19.5</b>	<b>17.0</b>	<b>0.333</b>
Employment <sup>d</sup>						
Unemployment of population aged 16 and older	<b>5.1</b>	<b>4.7</b>	<b>0.162</b>	<b>8.1</b>	<b>7.4</b>	<b>0.213</b>
Urbanicity <sup>a</sup>						
Population residing in urban areas	<b>43.8</b>	<b>39.2</b>	<b>0.15</b>	<b>45.6</b>	<b>40.3</b>	<b>0.166</b>

Data are mean percentages, and numbers highlighted in bold indicate non-negligible difference

<sup>a</sup> Data for 2000 were from Area Health Resource Files, and data for 2010 were from census 2010

<sup>b</sup> Data were from Small Area Income and Poverty Estimates

<sup>c</sup> Data were from Small Area Health Insurance Estimates for all ages in 2000 and for people <65 in 2010

<sup>d</sup> Data for 2000 were from Area Health Resource Files, and data for 2010 were from the American Community Survey 2005–2009

<sup>e</sup> Standardized difference

by 2010. Similarly, we found a higher Medicare managed care penetration rate in SEER areas than in non-SEER areas in 2000, whereas this difference disappeared by 2010. Finally, the SEER areas had higher percentages of population living in urban areas than the non-SEER areas, in 2000 and 2010.

As for the data quality, we found that case completeness ascertainment has meaningful difference between areas with SEER areas being better than non-SEER areas in every year (Table 3). The percentages of missing data in race were small (<4 %) and not different between SEER and non-SEER areas (data not shown). Overall, the average percentages of missing data in age and sex variables were almost zero in the majority of the registries (data not shown), and no difference was found between the two comparison areas.

## Discussion

This paper updates some important comparisons between SEER and non-SEER areas, in several dimensions, to assess the current generalizability of the SEER Registries. First, using USCS data we examined the representativeness of proportions of late-stage BC and CRC diagnoses, by cancer population subgroups. We found males and females with other race (persons who were not white, African-American, Asian, or Hispanic) were less likely to be diagnosed with late-stage CRC in the SEER versus non-SEER areas. On the other hand, Asian and Hispanic females aged <40 were more likely to be diagnosed at late-stage CRC in SEER areas than that in non-SEER areas. No difference was found for females with BC between the SEER and non-SEER comparison areas. We concluded that the SEER registry areas slightly over-represent Asian and Hispanic female subgroups with CRC, slightly under-represent people of other race with CRC, and are not different across the vast majority of comparisons. In addition, our findings using cancer cases in 2004 to 2009 for proportion of late-stage BC and CRC were similar to results from Henley et al. [7].

One finding worth noting is the lower proportion of late-stage CRC for several age groups of men and women in the other race or ethnicity group in SEER areas as compared to non-SEER areas. The relatively lower proportion of late-stage CRC in the SEER areas may be partially due to the fact that American Indians and Alaska Natives (AI/AN) were excluded from SEER areas in our study design. Given the current structure of the SEER program that includes only AI/AN from Alaska, Arizona, and Cherokee Tribal Jurisdictional Service Area of Oklahoma, it is not possible to assess the difference between SEER and non-SEER areas for ‘other’ racial or ethnic group. Future study is needed to better understand the disparities of cancer burden with this racial or ethnic group. To accomplish this, we need the USCS database covering the entire nation, which would not be apparent in studies using SEER registry data.

The important focus of this study was the comparison of population demographics and socioeconomic factors between SEER and non-SEER areas. In these comparisons, we found that SEER areas had a higher percentage of younger people and lower percentage of white population than non-SEER areas, which is consistent with the comparison by Warren et al. [2] using 1998 registry data with 1990 census data for the elderly US population. By contrast, we found that SEER areas had higher percentages of poor and a lower average employment rate, which is inconsistent with findings by Warren et al. [2] and Nattinger et al. [16] from earlier time periods. Those studies reported that SEER regions were more affluent and had lower unemployment than the rest of the nation. This inconsistency between our study and earlier studies may be explained by the expansion of SEER registries in 2000 to include Kentucky, Louisiana, New Jersey, and Greater California, as these states (except New Jersey) have higher than national average poverty rates [24]. Furthermore, after this expansion, SEER areas still have more people living in urban areas than the rest of the nation, similar to the previous findings by Warren et al. [2] and Nattinger et al. [16].

Our data also showed that Medicare managed care penetration rate was higher in SEER areas in 2000 but

**Table 3** Mean and standard deviation of case completeness between SEER and non-SEER areas by year

Year	SEER ( <i>N</i> = 15)	Non-SEER ( <i>N</i> = 40 <sup>a</sup> )	Standardized difference <sup>b</sup>
2008	105.8 (6.0)	101.6 (5.9)	<b>0.707</b>
2009	105.6 (6.6)	100.8 (5.1)	<b>0.874</b>
2010	103.1 (5.6)	100.0 (5.4)	<b>0.578</b>
2011	101.8 (5.7)	100.2 (5.4)	<b>0.308</b>
2012	103.6 (6.7)	99.7 (6.1)	<b>0.615</b>

<sup>a</sup> *N* is based on year 2008, and it varied by year as the following states were excluded in different years: Minnesota in 2008–2012, Arkansas in 2010, and Nevada in 2011–2012

<sup>b</sup> Number highlighted in bold are considered non-negligible difference



equivalent in 2010, compared to non-SEER areas. This may be explained by considerable expansion in the managed care program across the US after implementation of the Medicare Prescription Drug, Improvement, and Modernization Act in 2006 [25]. The Act renamed the Medicare + Choice program the Medicare Advantage (MA) program and made it much more attractive to seniors by adding prescription drug coverage to all MA plans. Specifically, Medicare managed care penetration rate increased from 15 % in 2000 to 24 % in 2010 and continued to increase [26]. Our data showed that non-SEER areas had higher increase in Medicare managed care penetration rate than SEER areas, with the average increase of 10.7 % and 8.8 %, respectively. Beginning in 2006, more isolated rural areas of the USA had access to a Medicare managed care plan for the first time. Prior to this, Medicare managed care penetration was predominately an urban phenomenon. Because the non-SEER areas are less urban/more rural than the SEER areas, these non-SEER areas saw a greater expansion in Medicare managed care plans than the more urban SEER areas, during the time of our study.

As regard to data quality issues, our data indicate that SEER areas tend to have more complete data than the non-SEER areas in terms of case completeness. This is not surprising as SEER registries were originally selected for program inclusion because of their relatively higher data quality. However, the conclusion that non-SEER areas have lower data quality needs to be stated with caution. First, the measure of case completeness used in the quality assessment assumes the ratio of age-adjusted cancer incidence rate to age-adjusted mortality rate is constant across geographic areas within cancer sites, race, and gender [22]. In addition, the measure is based on two race groups, white and black, and ignores the population in all other racial or ethnic groups. Thus, the measure may not well reflect the true completeness of case ascertainment in both SEER and non-SEER areas. Although NAACCR uses several adjustments to make the measurement more accurate, other researchers have proposed a new method to evaluate the case completeness [27]. The new method, taking into account geographic heterogeneity and including more registry-specific data to predict expected incidence rates, showed an improved case completeness estimate than the method from NAACCR. Future research is needed to re-evaluate the differences in case completeness between SEER and non-SEER areas.

Although the USCS data cover the entire nation, our conclusions regarding the late-stage BC and CRC comparisons across SEER and non-SEER areas are based on only 46 of the 50 states (Kansas, Maryland, Minnesota, and Virginia were not included in our analyses). However, the state-level incidence rates of BC and CRC in these four

excluded states are similar to the national average [28], so it is unlikely that our findings will be significantly different from other analyses using all states and had these states' data been available.

In summary, using BC and CRC data our study demonstrated that SEER registries were not significantly different from non-SEER areas in terms of average age–race-specific proportions of late-stage BC or CRC, except for estimates of late-stage CRC for other race and young Asian and Hispanic women. Although case completeness was better in SEER areas than non-SEER areas, SEER areas had greater economic disadvantage and greater minority diversity among the population. Researchers using SEER data who attempt to generalize results to the entire nation need to be aware of these differences.

Our study also demonstrated several advantages to researchers from having available the entire USCS database in the RDCs. It allows researchers to study the entire cancer population of the US, with data quality comparable to the well-known SEER registries. The database which covers the entire nation also allows researchers to conduct geospatial analysis, and to examine local community or contextual predictors of late-stage diagnoses and other cancer outcomes. The database also allows better examination of racial and ethnic disparities in late-stage diagnoses of CRC or other cancers, which would not be possible by using SEER data alone. In addition, for rare cancers, USCS data have the advantage of providing a larger sample size than SEER data—about three times larger for BC and CRC. Larger sample sizes for rare cancers may allow for more rigorous analyses than are possible with SEER registry data alone. These rich opportunities for population-based research were not possible prior to the release of these RDC-enabled combined registry data.

**Acknowledgments** CDC's National Program of Cancer Registries contributed funds to cover the standard RDC fees for researchers conducting analyses under approved research projects. This work was also supported by a National Cancer Institute Grant (2R01CA126858). The content is solely the responsibility of the authors and does not necessarily represent the official views of Georgia State University, the University of North Carolina, the National Center for Health Statistics, the National Cancer Institute, or the National Institutes of Health.

## References

1. National Cancer Institute (2016) Surveillance, epidemiology, and end results program. <http://seer.cancer.gov/registries/>. Accessed 20 Feb 2016
2. Warren J, Klabunde C, Schrag D, Bach P, Riley G (2002) Overview of the SEER-medicare data: content, research applications, and generalizability to the united states elderly population. *Med Care* 40(8):IV-3–IV-18

3. Research Data Center (RDC) (2016) United States cancer statistics. <http://www.cdc.gov/rdc/b1datatype/dt131.htm>. Accessed 25 Feb 2016
4. Wingo P, Jamison P, Hiatt R, Weir H, Gargiullo P, Hutton M, Lee N, Hall I (2003) Building the infrastructure for nationwide cancer surveillance and control—a comparison between The National Program of Cancer Registries(NPCR) and The Surveillance, Epidemiology, and End Results (SEER) Program (United States). *Cancer Causes Control* 14:175–193
5. Center for Disease Control and Prevention (CDC), National Program of Cancer Registries (NPCR): NPCR Standards. <http://www.cdc.gov/cancer/npcr/standards.htm>. Access 6 June, 2016
6. Centers for Disease Control and Prevention (CDC) (2015). National Program of Cancer Registries, US Cancer Statistics. [www.cdc.gov/uscs](http://www.cdc.gov/uscs). Accessed 10 Sept 2015. CDC's NCHS Research Data Center, <http://www.cdc.gov/rdc>. Accessed 10 Sept 2015
7. Henley J, King J, German R, Richardson L, Plescia M (2010) Surveillance of screening-detected cancers (colon and rectum, breast, and Cervix)—United States, 2004–2006. In: *Proceedings of the morbidity and mortality weekly report*, vol 59, Nov 26, p SS-9
8. Young JL Jr, Roffers SD, Ries LAG, Fritz AG, Hurlbut AAH (eds) (2001) SEER summary staging manual—2000: codes and coding instructions. no. NIH Pub. No. 01-4969. National Cancer Institute, Bethesda
9. Mobley L, Kuo T, Watson L, Brown G (2012) Geographic disparities in late-stage cancer diagnosis: multilevel factors and spatial interactions. *Health Place* 18(5):978–990
10. Frey CM, McMillen MM, Cowan CD et al (1992) Representativeness of the surveillance, epidemiology, and end results program data: recent trends in cancer mortality rates. *J Natl Cancer Inst* 84(11):872–877
11. Merrill RM, Dearden KA (2004) How representative are the surveillance, epidemiology, and end results (SEER) program cancer data of the United States? *Cancer Causes Control* 15:1027–1034
12. Irby K, Anderson WF, Henson DE, Devesa SS (2006) Emerging and widening colorectal carcinoma disparities between Blacks and Whites in the United States (1975–2002). *Cancer Epidemiol Biomark Prev* 15:792–797
13. Jackson-Thompson J, Ahmed F, German RR, Lai SM, Friedman C (2006) Descriptive epidemiology of colorectal cancer in the United States, 1998–2001. *Cancer* 2006(107):1103–1111
14. Kuo TM, Mobley LR, Anselin L (2011) Geographic disparities in late-stage breast cancer diagnosis in California. *Health Place* 17:327–334
15. Rim SH, Seeff L, Ahmed F, King JB, Coughlin SS (2009) Colorectal cancer incidence in the United States, 1999–2004: an updated analysis of data from the National Program of Cancer Registries and the Surveillance, Epidemiology, and End Results Program. *Cancer* 115:1967–1976
16. Nattinger AB, McAuliffe TL, Schapira MM (1997) Generalizability of the surveillance, epidemiology, and end results registry population: factors relevant to epidemiologic and health care research. *J Clin Epidemiol* 50(8):939–945
17. Humes KR, Jones NA, Ramirez RR (2011) Overview of race and Hispanic origin: 2010. 2010 Census briefs. <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>. Accessed 25 Feb 2016; and Census 2000 PHC-T-1. Population by race and hispanic or latino origin for the United States: 1990 and 2000. <http://www.census.gov/population/www/cen2000/briefs/phc-t1/tables/tab04.pdf>. Accessed 25 Feb 2016
18. Howden LM, Meyer JA (2011). Age and sex composition: 2010. 2010 Census briefs. <http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf>. Accessed 25 Feb 2016
19. American Cancer Society, (2015). Cancer facts & Figures 2015, Atlanta: American Cancer Society. <http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf>. Accessed 25 Feb 2016
20. National Cancer Institute (2016) Surveillance, epidemiology, and end results program. Site recode ICD-O-3/WHO 2008 definition. [http://seer.cancer.gov/siterecode/icdo3\\_dwhohome/index.html](http://seer.cancer.gov/siterecode/icdo3_dwhohome/index.html). Accessed 20 Feb 2016
21. CINA: 2009–2013, Appendix D: data quality indicators by year and registry. [http://www.naacr.org/dataandpublications/cina\\_pubs.aspx](http://www.naacr.org/dataandpublications/cina_pubs.aspx). Access 6 June 2016
22. Hofferkamp J (2008). Standards for cancer registries volume III: standards for completeness, quality, analysis, management, security and confidentiality of data. Springfield (IL): North American Association of Central Cancer Registries, August 2008
23. Austin P (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 46(3):399–424. doi:10.1080/00273171.2011.56878
24. Dalaker J (2001). Poverty in the United States: 2000. Issued September 2001. <https://www.census.gov/prod/2001pubs/p60-214.pdf>. Accessed 25 Feb 2016
25. Megellas, M. (2006). Medicare modernization: the new prescription drug benefit and redesigned part B and part C. In: *Proceedings* (Baylor University. Medical Center) 19(1): 21–23
26. Neuman T, Casillas G, Jacobson G (2015) Medicare advantage and traditional medicare: is the balance tipping? Kaiser family foundation (KFF) issue brief: October 2015. <http://files.kff.org/attachment/issue-brief-medicare-advantage-and-traditional-medicare-is-the-balance-tipping>. Accessed 25 Feb 2016
27. Das B, Clegg LX, Feuer EJ, Pickle LW (2008) A new method to evaluate the completeness of case ascertainment by a cancer registry. *Cancer Causes Control* 19:505–525
28. National Cancer Institute (2016) State cancer profiles. <http://statecancerprofiles.cancer.gov/index.html>. Accessed 25 Feb 2016