Birthplace and survival among Asian women diagnosed with breast cancer in cancer registry data: the impact of selection bias

From SCARLETT S LIN, CYNTHIA D O’MALLEY, CHRISTINA A CLARKE and GEM M LE

Sir—The recent article by Pineda et al. on survival following breast cancer among Chinese, Japanese, and Filipinos living in three Surveillance, Epidemiology, and End Results (SEER) regions in the Western US (San Francisco/Oakland, Hawaii, Seattle/Puget Sound)1 addresses an important question regarding the association between birthplace and breast cancer survival. However, while SEER data are generally known to be of high quality, some variables such as American Joint Committee on Cancer (AJCC) staging, chemotherapy, and birthplace are problematic. Our concern focuses on the bias associated with the birthplace variable.

In the Pineda et al. paper, information on place of birth was used to define the Asian patients in the study as having been born in the US, or born in the countries of Asia.1 Birthplace information in the SEER registry is generally derived from the patient’s medical records or death certificate.2 This variable, as it exists in the SEER registry, is problematic because it has never been validated; thus, its accuracy is unknown. In addition, birthplace information is missing for a substantial proportion of patients in the registry and its level of completeness continues to worsen over the years.3 Despite its problems, this variable continues to be used in analyses of cancer incidence among Asian subgroups and their descendants;4–11 relationship of birthplace to tumour size and stage among Asians12 and Hispanics,13 and in the article by Pineda et al. on breast cancer survival.1

Given the methodological problems, study results using this variable are potentially misleading. We outline below some of our on-going work assessing the utility of the birthplace variable in the Greater Bay Area Cancer Registry (GBACR), which operates the San Francisco/Oakland and San Jose/Monterey SEER registries, and discuss the impact of our findings on the results from studies using the SEER birthplace variable.

In Pineda et al.’s paper, the proportion of Asian subjects missing birthplace information varied from 14.1% among Japanese to 26.6% among Chinese; these subject were excluded from the analyses.4 We found that in the GBACR, 67% of patients diagnosed in the period 1988–1996 had known birthplace information.2 This proportion varied across racial/ethnic groups, with Vietnamese and other Southeast Asians having the highest level of known birthplace (89%) and non-Hispanic whites having the lowest level (65%). We further found that within racial/ethnic groups, certain patient and hospital characteristics were associated with having known birthplace. Table 1 shows the adjusted odds ratios for selected patient and hospital characteristics on having known birthplace for Hispanics, Chinese, Japanese, Filipinos, and Vietnamese/Southeast Asians in the GBACR. Pineda et al. alluded to the fact that birthplace data were more likely to be known for deceased subjects; this is due to the additional report of information from death certificates. Among the various patient and hospital characteristics, the most significant factor associated with known birthplace was indeed vital status. However, while being deceased was associated with a greater likelihood of having recorded birthplace in all racial/ethnic groups, this effect was greater among Japanese and Hispanics than among Filipinos and Vietnamese/Southeast Asians. In addition, completeness of the birthplace variable in the GBACR was not random but varied with characteristics such as age, year of diagnosis, hospital teaching status, ownership status, and size.

Preliminary results from an on-going SEER-funded study comparing registry to self-reported birthplace in the GBACR revealed that the sensitivity and predictive value positives (PV+) of birthplace among specific Asian subgroups were quite high, indicating generally good validity. However, overall, Asians with unknown birthplace were less likely to be foreign-born than Asians with known birthplace (Table 2). This pattern was seen across all Asian subgroups except for Vietnamese, of whom the vast majority were foreign-born. Thus, exclusion of patients with unknown birthplace information would result in a biased study sample.

Since deceased patients are more likely to have complete birthplace information, the analyses conducted by Pineda et al.1 are likely to be affected by the selection bias such that the hazard rates of death due to breast cancer are underestimated for Chinese, Japanese, and Filipinos; the selection bias would result in biased hazard rate ratios comparing Asians to Caucasians (towards the null for Chinese and Filipinos, and away from the null for Japanese). In addition, inflation of the hazard rate of death due to breast cancer is greater for US-born Chinese, Japanese, and Filipinos than for their foreign-born counterparts, Pineda et al. showed unadjusted hazard rate ratios of 1.44, 1.35, and 0.83 comparing Asia-born to US-born for Chinese, Japanese, and Filipinos, respectively;1 these rate ratios can be considered to be biased away from the null for Chinese and Japanese and towards the null for Filipinos given the
selection bias. The degree of bias in the adjusted models would be related to the relationship among birthplace completeness and the independent and dependent variables in the models.

Finally, we wish to address several other issues in the paper by Pineda et al. First, it would have been more informative to see adjusted hazard rate ratios comparing foreign-born Chinese, Japanese, Filipinos to their US-born counterparts, separately for stage and for treatment. Understanding what proportions of the survival deficits are due to stage or to treatment can help us to better guide efforts towards improving early detection or towards encouraging women and their physicians to ensure that the appropriate treatment is given. Second, while it is true that SEER does not directly collect information on socioeconomic status (SES), the individual registries do collect information on place of residence. The geographical location can be assigned a census unit, such as a census tract, and information about the tract, such as the percentage of high school graduates, median household income, and percentage below poverty level, can be obtained from the census. Thus, neighbourhood-level SES information is certainly available for patients in the SEER database.

The SEER data provide a rich, population-based resource for research. The data reflect diverse populations and the continued data collection for over 25 years provides an unparalleled opportunity to examine trends in cancer incidence and survival. As with all datasets, the nuances associated with the collection of the variables, as well as changes over time in collection procedures and definitions, are critical to analysis.

Acknowledgements
Cancer data were collected by the Northern California Cancer Center under contract N01-CN-65107 with the National Cancer Institute, National Institutes of Health, and with support of the California Cancer Registry, a project of the Cancer Surveillance Section, California Department of Health Services, under subcontract 050M-8701/8-S1522 with the Public Health Institute. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services or the California Department of Health Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government or state of California.

References
As we note in the Discussion section of our paper, completeness did not use Caucasian women as a reference group because this of place of birth varied by both ethnicity and vital status, thus we analyses that used the birthplace variable. because of missing birthplace was meant to apply only to the affecting missing birthplace in SEER data. As stated in our Methods section which indicated the proportion of women excluded those with missing birthplace. The sentence in our Methods paper, we agree that selection bias due to missing place of birth might have affected our results on Asian breast cancer survival in the US by place of birth.2
Our results comparing survival among Chinese, Japanese, Filipino, and Caucasian women (Tables 1 and 2) did not exclude those with missing birthplace. The sentence in our Methods section which indicated the proportion of women excluded because of missing birthplace was meant to apply only to the analyses that used the birthplace variable.
Our results on survival by place of birth within each of the Asian ethnic groups did exclude those with missing birthplace. As we note in the Discussion section of our paper, completeness of place of birth varied by both ethnicity and vital status, thus we did not use Caucasian women as a reference group because this would be subject to selection bias. We also noted that we could not assess whether within each ethnic group there would be selection bias.

References

Letters to the Editor 513

Author’s response

From EMILY WHITE and MICHELE PINEDA
Sir—we appreciate the new information provided in the letter from Lin et al. on the accuracy of place of birth and factors affecting missing place of birth in SEER data.1 As stated in our paper, we agree that selection bias due to missing place of birth might have affected our results on Asian breast cancer survival in the US by place of birth.2

Our results comparing survival among Chinese, Japanese, Filipino, and Caucasian women (Tables 1 and 2) did not exclude those with missing birthplace. The sentence in our Methods section which indicated the proportion of women excluded because of missing birthplace was meant to apply only to the analyses that used the birthplace variable.

Our results on survival by place of birth within each of the Asian ethnic groups did exclude those with missing birthplace. As we note in the Discussion section of our paper, completeness of place of birth varied by both ethnicity and vital status, thus we did not use Caucasian women as a reference group because this would be subject to selection bias. We also noted that we could not assess whether within each ethnic group there would be selection bias. The new data presented in the letter indicate that within each ethnic group, those with missing information are more likely to be alive and US born. We agree that this would inflate the hazard rate of death for US born Asian women; thus each of our rate ratios would be underestimates (e.g. biased toward the null for Chinese women rather than away from the null as stated in the letter).

In terms of the other concerns expressed in the letter, after adjustment of demographic factors, there were no differences in survival by birthplace. Further adjustment for stage and/or treatment had little effect and the results remained non-significant. We cannot comment on adjustment for SES as we do not have that information available.

Authors’ response to Commentaries on Outcomes in Occupational Cancer Epidemiology

From GERARD MH SWAEN and LUDOVIC GPM VAN AMELSVOORT
Sir—we thank Professors Bracken1 and Savitz2 for their positive and stimulating commentaries on our article on false positive and true positive outcomes in occupational cancer epidemiology.3 Bracken rightly puts the finger on one of the key complicating elements of our analysis, the lack of a true gold standard. Indeed as Bracken phrases it: our study of false positive and true positive outcomes in itself may very well be a false positive study. The lack of a true gold standard implies that our distinction between false positive and true positive outcomes contains some degree, hopefully only marginal, of subjectivity.


Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

References

Authors’ response to Commentaries on Outcomes in Occupational Cancer Epidemiologya

From GERARD MH SWAEN and LUDOVIC GPM VAN AMELSVOORT
Sir—we thank Professors Bracken1 and Savitz2 for their positive and stimulating commentaries on our article on false positive and true positive outcomes in occupational cancer epidemiology.3 Bracken rightly puts the finger on one of the key complicating elements of our analysis, the lack of a true gold standard. Indeed as Bracken phrases it: our study of false positive and true positive outcomes in itself may very well be a false positive study. The lack of a true gold standard implies that our distinction between false positive and true positive outcomes contains some degree, hopefully only marginal, of subjectivity.