# SUPPLEMENTAL METHODS

## Cohort construction, data pooling and variable harmonization

The CBCSC comprised six epidemiologic studies of breast cancer etiology and/or prognosis that were initiated in the 1990’s and early 2000’s. Prior to initiation of the current study, a one-year pilot study was performed during which prognostic factors related to body size, physical activity, and co-morbidities were identified that showed considerable variation in their distribution across racial/ethnic groups and that potentially could be harmonized – i.e., given similarly interpretable values -- across the CBCSC studies. In addition, linkage with geocoded patient records allowed us to identify institutional and neighborhood social and built environment factors that also showed racial/ethnic variation (1). The study objectives and activities of the CBCSC were organized into four individual projects, each focused on a specific set of potential prognostic factors: Project 1: contextual factors (i.e. those related to socioeconomic and man-made (“built”) physical attributes of an individual’s surroundings (1, 2)); Project 2: physical activity; Project 3: body size; and Project 4: co-morbidities. The three case-control studies contributed data to all four projects, whereas the cohort studies contributed data to a subset of the projects. The systematic assembly of data for the four projects was facilitated by a Cancer Registry Data Core (CRDC) and a Questionnaire Data Core (QDC) which were formed for this study. The CRDC centralized and streamline the ascertainment of cancer registry variables from the California Cancer Registry (CCR), a state-mandated population-based cancer registry that is part of the SEER program. Through its regional registries, the CCR routinely collects patient data from medical records on age at diagnosis, sex, race/ethnicity, marital status, birthplace, and tumor characteristics (American Joint Committee on Cancer (AJCC) stage, tumor size, grade, nodal involvement, histology, estrogen receptor (ER) and progesterone receptor (PR) status, laterality, HER2 (although missing for most cases prior to 2004, and thus not further considered)), first course of treatment (extent of surgical resection, chemotherapy, radiation), and vital status (including cause of death for the deceased) through hospital follow-up and linkages to vital statistics, death records, and other databases. Given that multiple tumors of an individual are captured, data on cancer(s) prior to and subsequent to the qualifying tumors for CBCSC eligibility are also available insofar as the women remain a resident of California.

Each participating study provided either unique CCR identification numbers or personal patient identifiers to enable linkage to the CCR database. The CRDC consolidated and created a dataset of clinical (tumor characteristics and treatment data), survival, and census block group SES variables. Missing clinical data from the CCR, such as chemotherapy, were not supplemented with data available from contributing study sources, given that these data were not collected systematically by all studies. Vital status as of December 31, 2009, the end of follow-up period of this study, was used to ensure, to the extent possible, that all cases in the pooled analysis had comparable opportunity for follow-up. Breast cancer-specific deaths were derived from the underlying cause of death on the death certificate based on ICD-9 (174-175) or ICD-10 (C50) codes. Patients’ addresses at diagnosis are routinely geocoded by the CCR to a coordinate (latitude, longitude). Neighborhood SES at the block group level was assigned to cases with at least a zip code+4 digit postal extension, using a previously developed index that incorporates US census data on education, occupation, unemployment, household income, poverty, rent, and house values (3). Cases diagnosed prior to 1996 were assigned to the SES measure developed with the 1990 census data, and those diagnosed in 1996 or later were assigned to the measure using 2000 census data.

The QDC harmonized a select set of demographic and lifestyle factors considered to be relevant covariates for adjustment in the four individual projects. Each of the six studies provided specific questionnaire variables that were then harmonized and merged into a common dataset. This pooled dataset included basic demographic variables (date of birth, race/ethnicity, education, birthplace, age at migration to the US if foreign-born, language of interview), and the major suggested breast cancer risk and prognostic factors, including pregnancies (number and outcome of pregnancies, ages at first and last birth), menopause (type of menopause, age at menopause), family history of breast cancer (number of affected first-degree relatives, age at diagnosis of the affected relatives), smoking pattern (never/former/ current, number of cigarettes smoked per day) and alcohol consumption before breast cancer diagnosis.

For some variables (race/ethnicity, birthplace, education, parity, age at first birth, time since last birth, family history of breast cancer, age at diagnosis of affected relatives), the questionnaire response categories were similar across the studies although the definitions were not always identical. For example, although all studies collected information on family history of breast cancer in first-degree female relatives, several studies asked about cancer history in full siblings (LACE, CTS, MEC), whereas two studies collected data separately for both full- and half-sisters (AABCS, CARE), and this information was not specified in another study (SFBCS). We left these data as they were originally coded, given that it is not possible to re-define the categories used by the contributing studies. The differences in classification caused by these inconsistencies are likely to be minimal. Other variables such as alcohol consumption and smoking (cigarettes per day among former and current smokers) were collected as continuous variables in some studies (AABCS, CARE, SFBCS, LACE) but as categorical variables in other studies (CTS, MEC); the midpoint of a category was assigned for the latter group. The questions that were used to assess menopausal status varied across the six studies. However, because each of the studies had carefully developed its own algorithm to determine menopausal status, we relied on each study’s original classification of menopausal status.

Of the initial 12,787 breast cancer case participants from the six studies submitted to the CRDC, 577 cases were excluded for various reasons. For the 492 cases who participated in more than one study, resulting in 506 duplicate records, data were included from case-control studies first and then from the cohort studies in the order of LACE, CTS, and MEC. The final analytic dataset included 12,210 breast cancer cases. For a small group of cases (n=126) with multiple breast tumors diagnosed on the same day, we designated the study-qualifying tumor based on a combination of stage, grade, and histology, and considered the worst prognosis tumor as the qualifying tumor.

Additional details of the construction of the CBCSC analysis cohort and dataset and the variables used can be found in (4-9)

## Baseline model

A common analytic approach was developed to facilitate the evaluation of the degree to which the residual race/ethnicity differences in overall and breast cancer-specific mortality could be explained by racial/ethnic differences in social and built environment and neighborhood factors, physical activity, body size, and co-morbidity variables after controlling for important tumor and lifestyle factors. We first developed a main effects stratified Cox regression model to estimate hazard ratios (HRs) and associated 95% confidence intervals (CIs) by reverse stepwise selection. The starting model included ‘study’ as a stratification factor and subject, tumor, and above-mentioned lifestyle and contextual factors available from all studies. Age at diagnosis was included as a continuous variable on both the natural scale and the log scale in order to appropriately account for the ordered effect of age and allow for non-linearity in its effect in the Cox model. Two time scales, time from diagnosis and attained age (10, 11) were investigated in the development of the final baseline models for overall and breast cancer-specific mortality. All variables, except for age at diagnosis were subject to removal using backward stepwise regression. The order of removal was determined by the more significant (between the two time scales) of the Cox partial likelihood ratio test for that variable, and removal continued until all remaining variables had a likelihood ratio p-value of <0.20, resulting in the final variable set for the baseline model(6). No interactions were included in the model. Women in the case-control studies (AABCS, CARE, SFBCS) and the prospective survivor cohort (LACE) survived after diagnosis until the time of data collection; thus their follow-up was left censored since women who died or were lost to follow-up before data collection by the parent study were not included in this study. Women in these four studies were admitted to the risk set at the time of data collection rather than at the time of diagnosis. Women in the prospective population cohorts (CTS, MEC) were included if their breast cancer diagnosis occurred during the study follow-up period. Women were followed until death; for breast cancer-specific death, we censored women who died of other causes on their dates of death.

## Disparity test

Tests of significance of a single variable of interest or sets of variables of interest when added to a reference Cox model were based on the Cox partial likelihood ratio test. Of note, the likelihood ratio comparison of a reference model to the reference model plus additional variables is not a direct test of only the change in the above that are used to compute the disparity measure *D.* Thus an additional likelihood ratio-based approximate test of the change in disparity was devised and performed. This test is referred to as the disparity .

We used an hypothesis test of the change in the log hazard ratios resulting from adding variables to a Cox regression model as an approximate test for the change in disparity measure *D*. Consider a Cox regression model with linear predictor where *X* is the vector of indicators for racial/ethnic group, *Y* is a vector of additional covariates for a baseline model, and *Z* is a vector of covariates whose influence on disparity we are interested in testing. , , and are the corresponding parameter vectors. When *Z* is added to or removed from the model, the change in log likelihood results from the change in the parameters (from a zero-vector to a non-zero vector and *vice versa*), as well as to changes in the value of , from which our disparity measure *D* is computed (see Statistical Methods), and to changes in the other parameters, , in the model. We wanted to gauge how much of this change in log likelihood is due only to the change in .

Let be the maximum partial likelihood estimate of the racial/ethnic parameters under the Cox regression model with linear predictor . Let be the maximum partial likelihood estimate of the racial/ethnic parameters, and be the maximum log partial likelihood value, for the Cox regression model with linear predictor (Model 1). Also, let be the maximum log partial likelihood value for a model with linear predictor (Model 0), i.e., the log partial likelihood value obtained by maximizing with respect to and , but with the constraint that is fixed at the value . Then , which we refer to as the disparity , is a chi-square test of the null hypothesis and represents in a sense that portion of the likelihood ratio chi-square of the null hypothesis that is due to the change in the value of the parameter resulting from adding covariates *Z* to the model.

Note that this test is not strictly a test of the difference in disparity measures *D* that is obtained under Model 1 and Model 0, as different values of do not necessarily result in different values of *D*. However, lack of statistical evidence of a difference between Model 1 and Model 0 implies lack of statistical evidence of a difference in the corresponding *D*s.


# References

1. Gomez SL, Glaser SL, McClure LA, Shema SJ, Kealey M, Keegan TH, et al. The California Neighborhoods Data System: a new resource for examining the impact of neighborhood characteristics on cancer incidence and outcomes in populations. Cancer Causes Control. 2011;22:631-47.

2. Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the built environment for physical activity: state of the science. Am J Prev Med. 2009;36:S99-123 e12.

3. Yost K, Perkins C, Cohen R, Morris C, Wright W. Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. Cancer Causes Control. 2001;12:703-11.

4. Kwan ML, John EM, Caan BJ, Lee VS, Bernstein L, Cheng I, et al. Obesity and mortality after breast cancer by race/ethnicity: The California Breast Cancer Survivorship Consortium. Am J Epidemiol. 2014;179:95-111.

5. Lu Y, John EM, Sullivan-Halley J, Vigen C, Gomez SL, Kwan ML, et al. History of Recreational Physical Activity and Survival After Breast Cancer: The California Breast Cancer Survivorship Consortium. Am J Epidemiol. 2015;181:944-55.

6. Wu AH, Gomez SL, Vigen C, Kwan ML, Keegan TH, Lu Y, et al. The California Breast Cancer Survivorship Consortium (CBCSC): prognostic factors associated with racial/ethnic differences in breast cancer survival. Cancer Causes Control. 2013;24:1821-36.

7. Cheng I, Shariff-Marco S, Koo J, Monroe KR, Yang J, John EM, et al. Contribution of the Neighborhood Environment and Obesity to Breast Cancer Survival: The California Breast Cancer Survivorship Consortium. Biomarkers and Prevention. 2015.

8. Wu AH, Kurian AW, Kwan ML, John EM, Lu Y, Keegan TH, et al. Diabetes and other comorbidities in breast cancer survival by race/ethnicity: the California Breast Cancer Survivorship Consortium (CBCSC). Biomarkers and Prevention. 2015;24:361-8.

9. Shariff-Marco S, Yang J, John EM, Kurian AW, Cheng I, Leung R, et al. Intersection of Race/Ethnicity and Socioeconomic Status in Mortality After Breast Cancer. J Community Health. 2015.

10. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. Am J Epidemiol. 1997;145:72-80.

11. Pencina MJ, Larson MG, D'Agostino RB. Choice of time scale and its effect on significance of predictors in longitudinal studies. Stat Med. 2007;26:1343-59.