Cancer Causes Control (2008) 19:1305–1318
DOI 10.1007/s10552-008-9202-1

ORIGINAL PAPER

Race/ethnicity and breast cancer estrogen receptor status: impact of class, missing data, and modeling assumptions

Nancy Krieger · Jarvis T. Chen · James H. Ware · Afamia Kaddour

Received: 15 February 2008 / Accepted: 25 June 2008 / Published online: 14 August 2008

Abstract
Objective To test whether reported associations between race/ethnicity and breast cancer estrogen receptor (ER) status are inflated due to missing ER data, lack of socioeconomic data, and use of the odds ratio (OR) rather than the prevalence ratio (PR).
Methods We geocoded and added census tract socioeconomic data to all cases of primary invasive breast cancer ($n = 42{,}420$) among women diagnosed between 1998 and 2002 in two California cancer registries (San Francisco Bay Area; Los Angeles County) and analyzed the data using log binomial regression.
Results Adjusting for socioeconomic position and tumor characteristics, in models using the imputed data, reduced the PR for the black versus white excess risk of being ER− from 1.76 (95% CI: 1.66, 1.86; adjusted for age and catchment area) to 1.47 (95% CI: 1.38, 1.56). The latter parameter estimate was 16% greater (i.e., 1.56) in models excluding women with missing ER data, and was 43% greater when estimated using the OR (i.e., 1.82).
Conclusion(s) Studies on race/ethnicity and ER status that fail to account for missing data and socioeconomic data and report the OR are likely to yield inflated estimates of racial/ethnic disparities in ER status.

Keywords Breast cancer estrogen receptor status · Health disparities · Epidemiology · Race/ethnicity · Socioeconomic position · Poverty · Black · Hispanic · Asian and Pacific Islander

This investigation was funded by NIH grant 1 R03 CA125839-01, issued by the National Cancer Institute.

N. Krieger (✉)
Department of Society, Human Development and Health, Harvard School of Public Health, 677 Huntington Avenue, Kresge 717, Boston, MA 02115, USA
e-mail: nkrieger@hsph.harvard.edu

J. T. Chen
Department of Society, Human Development and Health, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

J. H. Ware
Dean for Academic Affairs and Development of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

A. Kaddour
Department of Global Health and Population, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

An apparent scientific consensus holds that US racial/ethnic groups intrinsically have disparate distributions of breast cancer estrogen receptor (ER) status, with white women purported to have the highest prevalence—and black women the lowest—of ER-positive (ER+) tumors [1–4]. Nevertheless, studies on this topic are affected by several limitations. Among US epidemiologic investigations designed to explore associations between race/ethnicity and ER status, virtually all of the 19 studies reporting positive associations (usually crude): (a) relied on medical records for ER status data, (b) had a high percentage of missing data on ER status (upwards of 10–20% or more), with the data most likely to be missing for women of color (largely if not solely comprised of black women), and (c) included little or no socioeconomic data [3, 5–20]. By contrast, the 9 studies reporting no association between race/ethnicity and ER status typically: (a) relied on laboratory assays performed specifically for the study, (b) had little or no missing data on ER status

(0–3%), and (c) controlled for socioeconomic position, and also reported associations between socioeconomic position and ER status [21–29]. Thus, significant associations between race/ethnicity and breast cancer ER status (chiefly comparing US black to white women) derive chiefly from studies with a relatively high degree of missing data on ER status and no socioeconomic data.

If the data on ER status were truly missing completely at random, and if ER status were unrelated to socioeconomic position, then estimates of racial/ethnic disparities in breast cancer ER status in these prior studies would be unbiased [30–32]. Indicating that concerns about bias may be warranted, however, evidence suggests: (a) ER status is more frequently missing among women of color and/or less affluent women [3, 13–15, 26, 33], most likely because of inadequacies of medical care [1, 33, 34], and (b) the major known risk factors for ER status—both those affecting endogenous hormone levels (e.g., hormone therapy, nulliparity, late age at first pregnancy, postmenopausal obesity) and those reflecting quality of medical care (e.g., stage of diagnosis, tumor size)—are strongly associated with socioeconomic position, within and across diverse racial/ethnic groups [1, 26, 33–37]. Since ER status is a key tumor biomarker relevant to both breast cancer treatment and survival [1–4, 33–35], it is thus important to gauge how taking into account issues of missing data and confounding affect estimates of racial/ethnic disparities in ER status.

Recognizing deficiencies in extant research on ER status, one major review has recommended use of better and more consistent assays for ER status [1]. Also germane are longstanding debates over how racial/ethnic disparities in health are conceptualized: as embodied biological expressions of social inequality that are socially determined, versus biological consequences of intrinsic "racial" (usually meaning "genetic") differences [38–40]. Cognizant of the implications of these debates for research on racial/ethnic health disparities, another review has called for research on how socioeconomic position "contributes to the stage, age at diagnosis, and biology of breast carcinoma." [34, p. 1995]. Accordingly, guided by the ecosocial theory of disease distribution[41–43] and its concern with both societal determinants of health inequities and biased assumptions affecting health research, we sought to examine how estimates and explanations of racial/ethnic inequities in ER status might be biased by missing ER data and omission of socioeconomic data.

The specific a priori hypothesis, we sought to test was that estimates of racial/ethnic inequities in ER status would be attenuated by: (1) using appropriate methods to address issues of missing data, and (2) controlling for socioeconomic position. Deciding on the appropriate analytic methods for testing our hypotheses, moreover, led us to recognize a previously unremarked characteristic of most research on breast cancer ER status and race/ethnicity: their virtually exclusive reliance on the odds ratio [3, 6, 7, 9, 12–14, 17–20, 22–24, 26, 27], at times explicitly interpreted as a relative risk [12, 19]. Yet, the prevalence of ER+ (the most commonly analyzed outcome; prevalence $\approx$ 75%) and ER− (prevalence $\approx$ 25%) both substantially exceed the "rare" disease condition (<10%) required for the odds ratio to provide a valid estimate of the risk ratio [44–46]. Our third question accordingly concerned whether interpretation of results would be influenced by choice of parameter estimate, i.e., the odds ratio (OR) versus prevalence ratio (PR).

Materials and methods

Study population

The study base consisted of the population residing, between 1998 and 2002, in the catchment area of two well-established population-based cancer registries: (1) the Northern California Cancer Center's (NCCC) San Francisco/Oakland SEER cancer registry, encompassing FIVE counties (Alameda, Contra Costa, Marin, San Francisco, and San Mateo) [47], and (2) the Los Angeles Cancer Surveillance Program (LA CSP), encompassing Los Angeles County [48]. We chose these registries for three reasons: (1) demographically, their catchment areas have substantial heterogeneity with respect to socioeconomic position and were sufficiently large with enough racial/ethnic diversity to permit meaningful sub-analyses among white, black, Asian and Pacific Islander, and Hispanic populations; (2) the regions they cover are relatively high incidence areas for breast cancer, with rates on average exceeding or equaling those of all SEER registries combined; and (3) they rank highly for the completeness (estimated at $\geq$98%), timeliness, and accuracy of registering cancer cases [47–49]. All analyses performed for this study were approved by the Harvard School of Public Health Human Subjects Committee and the Institutional Review Boards of both cancer registries. Since we were provided only de-identified records for a secondary data analysis, we were not required to obtain informed consent from the women included in the cancer registries.

Breast cancer cases

From this study base, we included all cases of primary invasive breast cancer among women recorded by the two cancer registries as being diagnosed between 1 January 1998 and 31 December 2002 (n = 42,240). We obtained data from the cancer registries on: age at diagnosis,

race/ethnicity, estrogen receptor status, tumor stage, tumor size, histologic type, and residential address at time of diagnosis.

All patient data were obtained from medical charts and it is unknown whether their racial/ethnic data were based on self-report or observer-report [47–49]. The racial/ethnic categories employed by US cancer registries correspond to those used in the US census, with these categories defined by the US Office of Management and Budget as "social-political constructs and should not be interpreted as being scientific or anthropological in nature."[50] Using the cancer registry racial/ethnic categories, we delineated the following mutually exclusive groups: white non-Hispanic ($n = 26{,}491$), black non-Hispanic ($n = 4{,}102$), Asian and Pacific Islander non-Hispanic ($n = 4{,}970$), American Indian non-Hispanic ($n = 38$), "other race" non-Hispanic ($n = 356$), and Hispanic ($n = 4{,}961$). Research on racial/ethnic misclassification of cancer registry and hospital records in California [51–54] and in the US nationally [55] indicates that while the sensitivity and specificity of racial/ethnic classification for the white and black population is reasonably high (in excess of 95%), it is somewhat lower for other racial/ethnic groups. In our racial/ethnic-specific analyses, we do not include data on the American Indian and "other race" non-Hispanic women since small numbers preclude meaningful analyses of these data.

In the cancer registry records [47, 48], ER status was defined as: (a) *positive*: test done and results were positive; (b) *negative*: test done and results negative; and (c) *unknown*: "test not done (includes cases diagnosed at autopsy)"; "test done, results borderline or undetermined whether positive or negative"; "test ordered, results not in the chart"; or "unknown if test done or ordered; no information (includes death-certificate-only cases)." Among cases missing ER status, the most common category was "unknown if test done or ordered" (78%) followed by "test not done" (16%). No data were available on reproductive history or hormone therapy use, precluding analysis of ER data in relation to these variables.

Socioeconomic measures

We geocoded the breast cancer cases included in this study using a commercial geocoding company whose accuracy we previously had tested and found to be high (96%) [49, 56]. We accepted only results geocoded to high precision (based on either exact street address or ZIP + 4 code; the latter is an area typically the size of one city block). We were able to geocode fully 97% of our cases with high precision to their census tract (CT) geocodes, the geographic level chosen because, as shown by results of our prior *Public Health Disparities Geocoding Project* [56–59], the census tract provided maximal geocoding and linkage to area-based socioeconomic data (compared to block group and ZIP Code data) and consistently detected expected socioeconomic gradients in health across a wide range of health outcomes.

We selected and constructed our CT area-based socioeconomic measures (ABSMs) based on theoretical considerations and methods described in detail in the publications of the *Public Health Disparities Geocoding Project* [56–59]. ABSMs generated and available pertain to CT poverty, income, occupation, education, and several deprivation indices. For analyses we previously conducted of socioeconomic gradients in breast cancer incidence [49], we found results overall were robust to choice of ABSM and that the ABSM that most informatively delineated the socioeconomic gradient was a new composite variable we created combining data on poverty and high income (defined as $\geq$4 times the US median household income, and calculated from the categorical income distribution by interpolation, assuming a Pareto distribution within the income category) [49]. The composite measure employed five mutually exclusive categories: (1) <5% below poverty and $\geq$10% high income; (2)<5% below poverty and <10% high income; (3) 5.0–9.9% below poverty; (4) 10.0–19.9% below poverty; and (5) $\geq$20% below poverty (the federal definition of a poverty area [60]).

Additionally, because of the strong association documented between educational level and ER status [26], we also employed an ABSM pertaining to the proportion of adults age 25 and older who had completed four or more years of college education. The pairwise correlations between the three ABSMs used to create these measures (percent below poverty, percent high income, percent college graduates) were all modest ($r < 0.4$), indicating they were not collinear. The proportion of the study catchment population living in CTs for which the ABSM data were missing was small (0.0–0.3%) and did not vary by race/ethnicity.

Statistical analyses

Our analytic plan involved four steps. First, we determined the univariate distribution, within our study population, overall and by race/ethnicity, of both the study outcome (ER status) and the specified covariates (age, socioeconomic position, tumor stage, tumor size, histologic type), as well as each variable's extent of missingness. We then created our analytic data set by excluding the small number of women ($n = 496$) missing data, singly or jointly, on the composite ABSM ($n = 12$; 0.03%), the college graduate ABSM ($n = 2$; 0.005%), the registry "race" variable ($n = 349$; 0.85%), and the registry "Hispanic" variable ($n = 410$; 1%). We opted not to impute these variables for two reasons: (1) the small number missing, and (2) maintaining comparability to the prior

literature on racial/ethnic disparities in ER status, which included women only of known race/ethnicity.

Second, for each variable, using the relevant referent group, we calculated the crude OR and PR for being: (1) ER+ versus ER−, (2) ER− versus ER+, and (3) ER status unknown versus ER status known, among cases with completely observed data (prior to imputation), in order to assess the extent to which estimates of racial/ethnic disparities would be affected by these analytic choices. As noted previously, most prior research, has focused on estimating the OR for being ER+ [3, 5, 7, 10, 11, 14, 16, 17, 19, 21, 23–28], Arguably, however risk of being ER− might be the more appropriate parameter, given that ER− is the more adverse outcome and also the rarer outcome (and hence less likely to result in the OR providing a biased estimate of the risk ratio [44–46]).

As our third step, we then employed multiple imputations to address potential limitations arising from analyzing only observations with fully observed data [30, 31]. The variables we imputed were: estrogen receptor status (21% missing), tumor stage (2% missing), and tumor size (7% missing). As our data set included the most important known predictors of both ER missingness (e.g., race/ethnicity and socioeconomic position) and ER status (e.g., sociodemographic and tumor characteristics), it is reasonable to posit that our use of multiple imputation was justified, given the key Missing At Random (MAR) criterion, whereby the probability of missingness depends only on variables that are observed [30, 31]. We conducted the imputation using the Amelia II program [61] to create 20 multiply imputed data sets and combined results using the SAS PROC MIANALYZE procedure.

Fourth, using the data set with imputed values, and informed by the results of the preceding analyses, we built up models to assess the prevalence rate ratio of being ER− versus ER+ in relation to race/ethnicity and to socioeconomic position, independently and together, adjusting for relevant covariates (age, catchment area, tumor size, tumor stage, and histologic type). For these models, we used log binomial regression, an analytic approach specifically developed for conditions in which the "odds ratio is not a good approximation of the risk or prevalence ratio." [62] The parameter estimates from these models can be expressed as prevalence ratios [62–65]. In order to calculate the percent change in excess risk comparing two parameter estimates (e.g., PR1 vs. PR2), we used the formula: ((PR1-1)−(PR2-1))/(PR1-1). Due to unexpected catchment area differences in the prevalence of ER unknown tumors (higher in Los Angeles than in the San Francisco Bay Area, as shown in Table 1), we tested for interaction effects between race/ethnicity and catchment area for risk of ER status; finding none, we controlled for catchment area in the models. We conducted all analyses in SAS [66].

Results

Table 1 presents selected descriptive data on the study population distribution, by estrogen receptor (ER) status, on the distribution of tumor characteristics (stage, size, histologic type), socioeconomic position, and catchment area, overall and by race/ethnicity. Highlighting the strong association between race/ethnicity and socioeconomic position, a much higher proportion of the black non-Hispanic, Hispanic, and Asian and Pacific Islander non-Hispanic cases, compared to white non-Hispanic cases, i.e., 47.8%, 34.6%, and 16.1%, versus 7.6%, respectively, lived in impoverished census tracts (20+% below poverty).

Figure 1 visually depicts the patterning of ER status by socioeconomic position across racial/ethnic groups. Overall, 21.0% of the women were missing ER status, with missingness highest among the Hispanic and black non-Hispanic women (28.5% and 24.7%, respectively), followed by the Asian and Pacific Islander non-Hispanic women (22.1%), and least among the white non-Hispanic women (18.4%). As would be expected, estimates of the percent ER+ and ER− were higher when based only on cases with known ER status, since these estimates ignore the percent with ER unknown. For the women overall, the contrast was 79.2% ER+ and 20.8% ER− (known ER status) versus 62.6% ER+ and 16.5% (all cases, including the unknown). By race/ethnicity, these contrasts were: (a) white non-Hispanic: 82.8% ER+ and 17.2% ER− (known ER status) versus 67.5% ER+ and 14.0% ER− (all cases); (b) black non-Hispanic: 65.8% ER+ and 34.2% ER− (known ER status) versus 49.6% ER+ and 25.7% ER− (all cases); (c) Asian and Pacific Islander non-Hispanic: 77.2% ER+ and 22.8% ER− (known ER status) versus 60.1% ER+ and 17.8% ER− (all cases); and (d) Hispanic: 72.3% ER+ and 27.7% ER− (known ER status) versus 51.7% ER+ and 19.8% ER− (all cases).

Also as expected, the distribution of ER status (both known and unknown), in addition to differing by race/ethnicity, varied by age at diagnosis, tumor characteristics, and socioeconomic position (Table 1). Both ER− tumors and tumors missing data on ER status were most common, and ER+ least common, among the younger women, women diagnosed with regional and distant tumors and with ductal histologic type (ER− only) or "other" histologic type (especially if ER unknown), and women living in the more impoverished and less educated census tracts. For example, 24% of the women with ER status unknown and 19% with ER− tumors, versus 13% of the women with ER+ tumors, lived in census tracts with 20+% poverty.

Table 2 shows results for multivariable analyses regarding racial/ethnic disparities for risk of having ER status unknown, analyzed in relation to the PR. The excess risk of having ER status unknown among the women of color

Table 1 Estrogen receptor status distribution of invasive breast cancer cases by age, tumor characteristics, and area-based socioeconomic position, overall and by race/ethnicity: San Francisco Bay Area* and Los Angeles County, 1998–2002

Characteristic	Total			Estrogen receptor status**: N and distribution (column %)											
				White non-Hispanic			Black non-Hispanic			Asian and Pacific Islander non-Hispanic			Hispanic		
	ER+ $n = 25{,}366$	ER− $n = 6{,}676$	Unk $n = 8{,}505$	ER+ $n = 17{,}424$	ER− $n = 3{,}620$	Unk $n = 4{,}761$	ER+ $n = 1{,}935$	ER− $n = 1{,}005$	Unk $n = 965$	ER+ $n = 2{,}898$	ER− $n = 856$	Unk $n = 1{,}068$	ER+ $n = 3{,}070$	ER− $n = 1{,}178$	Unk $n = 1{,}693$
Age (yrs)															
<45	10.9	20.4	14.0	8.4	16.6	8.9	13.0	21.9	14.0	18.1	21.6	20.0	17.0	30.0	24.5
45–54	21.6	27.7	21.9	19.5	26.0	18.4	22.1	30.9	22.6	30.5	30.6	28.7	25.4	28.2	27.2
≥55	67.4	51.9	64.1	72.1	57.5	72.8	65.0	47.2	63.4	51.4	47.8	51.2	57.5	41.9	48.3
SEER stage															
local	63.9	55.4	58.9	66.3	57.0	61.2	57.8	53.1	52.4	62.7	55.8	62.7	55.6	52.7	53.6
regional	32.5	39.0	26.8	30.3	37.5	25.0	37.1	39.6	28.5	34.6	39.4	25.0	39.8	42.4	32.4
distant	3.1	5.0	7.3	3.0	4.9	7.0	4.5	6.6	8.8	2.1	4.4	6.6	4.3	3.9	7.6
unspecified	0.4	0.6	7.0	0.4	0.4	0.4	0.6	0.6	0.6	0.5	0.5	0.5	0.4	0.4	0.4
Tumor size															
Missing	7.5	12.3	20.9	6.7	11.7	20.3	11.3	14.1	23.9	8.0	10.7	19.8	9.2	13.3	21.6
<10 mm	19.4	11.4	19.8	20.7	13.1	23.0	14.7	7.4	13.1	18.5	12.4	21.2	15.8	9.3	13.9
10–19 mm	39.2	27.7	24.9	40.7	31.0	26.6	36.6	25.2	21.0	37.4	26.3	23.6	33.6	21.1	22.9
20–29 mm	19.6	23.5	16.4	18.8	22.2	15.4	19.7	26.0	16.5	21.0	24.6	17.8	22.6	24.5	18.0
≥30 mm	14.3	25.1	18.0	13.1	22.0	14.6	17.7	27.4	25.5	15.1	25.9	17.7	18.8	31.7	23.6
Histologic type															
Ductal	64.2	76.3	56.9	62.7	75.8	54.0	64.3	77.2	58.2	72.6	78.6	61.9	65.0	75.4	61.1
Lobular—subtype 1	9.7	2.4	6.6	10.8	3.3	8.6	7.5	1.2	3.3	5.7	0.8	3.5	8.8	1.6	4.7
Lobular—subtype 2	13.8	4.0	8.2	14.7	4.8	10.0	12.8	2.2	6.2	10.2	2.8	5.7	12.3	4.0	6.1
Other	12.3	17.4	28.3	11.8	16.2	27.4	15.4	19.4	32.2	11.5	17.8	28.9	13.9	19.0	28.2
Area-based socioeconomic position															
*Census tract (CT) Poverty + High Income****															
<5% below poverty and 10+% ≥high income	20.6	15.7	14.2	25.9	23.0	21.7	4.5	3.2	2.5	14.6	13.9	9.6	6.1	5.3	2.9
<5% below poverty and <10.0 high income	11.5	11.3	9.6	12.6	14.5	12.6	5.1	4.4	2.3	11.7	11.3	10.2	9.0	7.5	4.8
5.0–9.9% below poverty	32.4	29.2	26.5	35.0	34.1	32.3	15.7	16.4	12.8	35.3	32.9	29.4	25.1	22.4	16.3
10.0–19.9% below poverty	22.1	24.8	26.0	19.4	21.3	23.0	30.6	28.7	26.9	23.9	25.9	30.1	30.3	31.4	31.2
≥20% below poverty	13.4	18.9	23.7	7.0	6.9	10.4	44.2	47.4	55.4	14.5	15.9	20.5	29.5	33.4	44.7
CT % college graduates															
40–100%	17.6	25.7	30.1	9.5	11.7	14.6	50.0	51.4	61.2	14.4	17.9	21.9	45.7	52.3	61.0
25–39.9%	16.8	18.4	18.4	14.8	17.6	17.5	19.7	19.1	17.4	22.3	19.5	22.9	21.1	19.4	18.4

Table 1 continued

Characteristic	Total			Estrogen receptor status**: N and distribution (column %)											
				White non-Hispanic			Black non-Hispanic			Asian and Pacific Islander non-Hispanic			Hispanic		
	ER+ $n = 25,366$	ER− $n = 6,676$	Unk $n = 8,505$	ER+ $n = 17,424$	ER− $n = 3,620$	Unk $n = 4,761$	ER+ $n = 1,935$	ER− $n = 1,005$	Unk $n = 965$	ER+ $n = 2,898$	ER− $n = 856$	Unk $n = 1,068$	ER+ $n = 3,070$	ER− $n = 1,178$	Unk $n = 1,693$
15–24.9%	22.6	22.3	20.0	23.7	25.4	23.1	11.7	13.4	9.2	28.6	26.9	27.4	17.7	17.4	12.5
0–14.9%	43.1	33.6	31.5	52.1	45.3	44.7	18.7	16.0	12.1	34.7	35.7	27.7	15.5	11.0	8.2
Catchment area															
SF Bay Area*	39.5	35.3	23.1	42.0	39.1	29.3	32.4	33.4	17.7	45.8	42.5	22.8	23.9	19.7	8.3
Los Angeles County	60.5	64.7	76.9	58.0	60.9	70.7	67.6	66.6	82.3	54.2	57.5	77.2	76.1	80.3	91.7

* San Francisco (SF) Bay Area: Alameda, Contra Costa, Marin, San Francisco, and San Mateo counties

** Estrogen receptor status: ER+ = positive, ER− = negative, Unk = unknown

*** High income was defined as households with median family income ≥4 times US median family income, which was $167,776 for the 2000 census

compared to white women (Model 1) was strongly attenuated by adjusting for socioeconomic position (Model 2), with the effect of this adjustment greater than adjustment for tumor characteristics and catchment area (Model 3). For example, comparing the black non-Hispanic to the white non-Hispanic women, the 33% greater crude risk for ER status unknown (Model 1) was reduced to 4% and rendered statistically non-significant in models that controlled only for socioeconomic position (Model 2), and remained statistically non-significant, at 7%, in models adjusting for all included covariates (Model 4). Similar patterns were evident for the Hispanic and the Asian and Pacific Islander non-Hispanic women, albeit the reduction in excess risk by controlling for the socioeconomic and other covariates was not sufficient to render the difference statistically non-significant.

Next, Table 3 presents the multivariable analyses for racial/ethnic and socioeconomic disparities, separately and combined, for being ER− versus ER+, as measured using the PR and based on the imputed data. Racial/ethnic and socioeconomic disparities were evident in models adjusting solely for age and catchment area (Models 1 and 2), with risk of being ER−, respectively greatest among the black non-Hispanic compared to the white non-Hispanic women (Model 1: PR = 1.76; 95% CI: 1.66, 1.86), followed by the Hispanic women (Model 1: PR = 1.42; 95% CI: 1.34, 1.50) and the Asian and Pacific Islander non-Hispanic women (Model 1: PR = 1.19; 95% CI: 1.11, 1.26) and lowest among women living in CT with the highest versus lowest proportion of college graduates (Model 2: PR = 0.71; 95% CI: 0.66, 0.76). As shown by Model 3, adding socioeconomic data to Model 1 had nearly as great an impact on reducing the estimates of racial/ethnic disparities in ER status as did separately adjusting, in Model 4, for tumor characteristics. In the fully adjusted Model 5 (including data on socioeconomic position, tumor characteristics, age, and catchment area), both black non-Hispanic and Hispanic women (but not Asian and Pacific Islander non-Hispanic women) remained at elevated albeit lower risk of being ER− (PR = 1.47; 95% CI: 1.38, 1.56 and PR = 1.21; 95% CI: 1.14, 1.29, respectively), as did women who lived in the lowest compared to highest income census tracts (PR = 1.10; 95% CI: 1.00, 1.20); women who lived in the most compared to least educated census tracts were at lowest risk (PR = 0.85; 95% CI: 0.79, 0.91).

Table 4 compares findings for: (a) the OR versus PR as the parameter estimate, using the imputed data, and (b) the PR, using the observed versus imputed data. Adjusting for age, socioeconomic position, tumor characteristics, and catchment area (Model 1), the OR was 43% greater than the PR for the black non-Hispanic/white non-Hispanic comparison (1.82 vs. 1.47), 34% greater for the Hispanic/

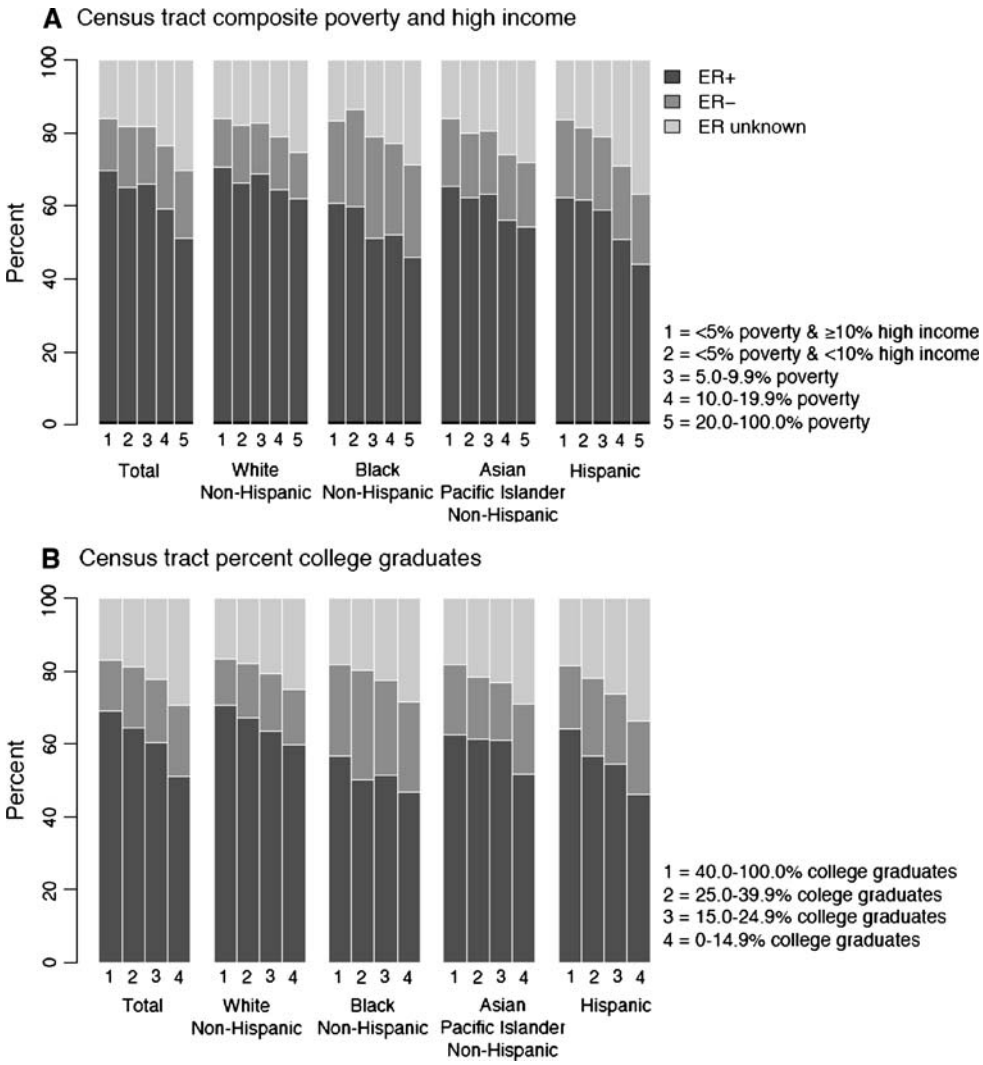

Fig. 1 Distribution of estrogen receptor (ER) status (positive, negative, unknown) among women with primary invasive breast tumors by race/ethnicity and two area-based measures of socioeconomic position: (**a**) census tract poverty/high income composite measure and (**B**) census tract percent of college graduates, San Francisco Bay Area and Los Angeles County, 1998–2002

white non-Hispanic comparisons (1.32 vs. 1.21), and 36% greater for the Asian non-Hispanic/white non-Hispanic comparison (1.11 vs. 1.07). Adjusting for the same covariates in Model 2, the PR for being ER− versus ER+ was reduced, for analyses based on the imputed versus observed data, by 16% for both the black/white and Hispanic/white comparisons (1.56 vs. 1.47, and 1.25 vs. 1.21, respectively), and by 13% for the Asian/white comparisons 1.08 vs. 1.07).

Discussion

The central finding of our investigation of racial/ethnic disparities in breast cancer ER status is that estimates of the magnitude of these disparities are sensitive to inclusion of socioeconomic data and treatment of missing data for ER status, as well as choice of parameter estimate, i.e., the prevalence ratio versus the odds ratio. Not only was the racial/ethnic patterning of missing ER data driven chiefly by racial/ethnic socioeconomic disparities, but the observed crude racial/ethnic disparities in ER status were notably reduced by adjusting for socioeconomic position and, to a lesser extent, by using imputed data. In the case of black/white comparisons, in analyses based on the imputed data, the excess risk measured by the OR for being ER− versus ER+ in the fully adjusted model was 43% greater than for the PR, and it was 16% higher for the PR in analyses based on the observed versus imputed data. The net implication is that studies on race/ethnicity and ER status that neglect to include socioeconomic data and fail to account for missing data will yield inflated estimates of racial/ethnic disparities in ER status, a problem magnified by reporting the OR rather than the PR.

Study limitations

Before accepting this study's results, it is important to consider potential limitations affecting the study design and data analysis. First, since we were able to obtain only

Table 2 Multivariable analysis* of prevalence ratio (PR) for racial/ethnic disparities in missing estrogen receptor (ER) status, overall and adjusting for socio-demographic and tumor characteristics: primary invasive breast cancer cases among women, San Francisco Bay Area**, and Los Angeles County, 1998–2002

Characteristics	Model 1		Model 2		Model 3		Model 4	
	PR	(95% CI)	PR	(95% CI)	PR	(95% CI)	PR	95% CI
Race/ethnicity								
White non-Hispanic	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Black non-Hispanic	1.33	(1.23 ,1.42)	1.04	(0.96,1.12)	1.22	(1.14,1.31)	1.07	(0.99,1.15)
Asian and Pacific Islander non-Hispanic	1.22	(1.14 ,1.31)	1.15	(1.07,1.23)	1.22	(1.14,1.30)	1.18	(1.10,1.26)
Hispanic	1.56	(1.48 ,1.65)	1.27	(1.20,1.35)	1.37	(1.30,1.45)	1.22	(1.15,1.30)
ABSM								
<5% poverty and ≥10% high income			0.74	(0.67,0.81)			0.91	(0.83,1.00)
<5% poverty and <10% high income			0.77	(0.70,0.85)			0.94	(0.85,1.03)
5.0–9.9% poverty			0.75	(0.69,0.81)			0.85	(0.79,0.91)
10–19.9% poverty			0.88	(0.82,0.94)			0.92	(0.86,0.98)
≥20% poverty			1.00	(reference)			1.00	(reference)
CT % college graduates								
40–100%			0.88	(0.82,0.95)			0.89	(0.83,0.96)
25–39.9%			0.80	(0.74,0.86)			0.83	(0.77,0.89)
15–24.9%			0.76	(0.71,0.82)			0.81	(0.75,0.88)
0–14.9%			1.00	(reference)			1.00	(reference)
Catchment area								
Los Angeles					1.93	(1.82,2.04)	1.84	(1.73,1.95)
SF Bay Area**					1.00	(reference)	1.00	(reference)
Age								
0–44					1.04	(0.97,1.10)	1.04	(0.97,1.11)
45–54					1.00	(0.94,1.05)	1.00	(0.95,1.05)
≥ 55					1.00	(reference)	1.00	(reference)
Stage								
localized					1.00	(reference)	1.00	(reference)
regional					0.83	(0.79,0.88)	0.84	(0.79,0.88)
remote					1.26	(1.13,1.40)	1.24	(1.12,1.38)
Tumor size								
0–10 mm					1.00	(reference)	1.00	(reference)
10–20 mm					0.70	(0.66,0.75)	0.70	(0.66,0.74)
20–30 mm					0.80	(0.75,0.85)	0.79	(0.74,0.84)
≥30 mm					0.97	(0.91,1.04)	0.94	(0.88,1.01)
Histology								
Ductal					1.00	(reference)	1.00	(reference)
Lobular, Type 1					0.98	(0.90,1.06)	0.99	(0.91,1.08)
Lobular, Type 2					0.86	(0.79,0.92)	0.87	(0.80,0.94)
Other					1.42	(1.35,1.50)	1.42	(1.35,1.50)

* Parameter estimates for each covariate adjusted for all other covariates in the model (indicated by all covariates appearing in each column)

** San Francisco Bay Area = Alameda, Contra Costa, Marin, San Francisco and San Mateo counties

CI = confidence interval

data included in cancer registry records, we lacked information on several known risk factors for breast cancer ER status: use of hormone therapy, postmenopausal obesity, and reproductive history, including both nulliparity and late age at first pregnancy [1, 33–37]. Given that all of these risk factors, except for postmenopausal obesity, are more prevalent in the US among more affluent, more educated, and white women, compared to women of color and to more economically deprived and less educated women [3, 34, 37, 49], then presumably adjusting for these additional

Table 3 Multivariable analyses* of racial/ethnic and socioeconomic disparities in the prevalence ratio (PR) for being ER− versus ER+, based on the imputed data, for primary invasive breast cancer among women, San Francisco Bay Area** and Los Angeles County, 1998–2002

Characteristics	Model 1		Model 2		Model 3		Model 4		Model 5	
	Imputed		Imputed		Imputed		Imputed		Imputed	
	PR	(95% CI)	PR	(95% CI)	PR	(95% CI)	PR	(95% CI)	PR	(95% CI)
Race/ethnicity										
White non-Hispanic	1.00	(reference)			1.00	(reference)	1.00	(reference)	1.00	(reference)
Black non-Hispanic	1.76	(1.66,1.86)			1.62	(1.52,1.73)	1.51	(1.42,1.59)	1.47	(1.38,1.56)
Asian and Pacific Islander non-Hispanic	1.19	(1.11,1.26)			1.15	(1.08,1.23)	1.08	(1.01,1.15)	1.07	(1.00,1.14)
Hispanic	1.42	(1.34,1.50)			1.31	(1.23,1.39)	1.25	(1.19,1.33)	1.21	(1.14,1.29)
Age										
0–44	1.65	(1.57,1.73)	1.73	(1.64,1.82)	1.65	(1.57,1.73)	1.45	(1.38,1.53)	1.45	(1.38,1.53)
45–54	1.36	(1.29,1.42)	1.40	(1.33,1.47)	1.36	(1.30,1.43)	1.30	(1.24,1.36)	1.30	(1.24,1.37)
≥ 55	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Stage										
Localized							1.00	(reference)	1.00	(reference)
Regional							0.97	(0.93,1.02)	0.98	(0.93,1.02)
Remote							0.97	(0.89,1.05)	0.96	(0.89,1.05)
Tumor size (mm)										
0–10							1.00	(reference)	1.00	(reference)
10–20							1.09	(1.01,1.17)	1.08	(1.00,1.17)
20–30							1.49	(1.38,1.60)	1.48	(1.37,1.59)
≥30							1.84	(1.70,1.98)	1.83	(1.69,1.97)
Histology										
Ductal							1.00	(reference)	1.00	(reference)
Lobular, Type 1							0.33	(0.28,0.39)	0.33	(0.28,0.39)
Lobular, Type 2							0.37	(0.33,0.41)	0.37	(0.33,0.42)
Other							1.11	(1.06,1.17)	1.11	(1.06,1.17)
Catchment area										
SF Bay Area**	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Los Angeles	1.13	(1.08,1.18)	1.08	(1.04,1.14)	1.09	(1.04,1.15)	1.13	(1.08,1.18)	1.12	(1.07,1.17)
CT poverty/high income measure										
<5% poverty and ≥10% high income			0.89	(0.81,0.98)	1.00	(0.91,1.10)			1.10	(1.00,1.20)
<5% poverty and <10% high income			0.95	(0.87,1.03)	1.05	(0.97,1.14)			1.12	(1.04,1.22)
5.0–9.9% poverty			0.90	(0.84,0.97)	1.00	(0.92,1.08)			1.06	(0.98,1.14)
10–19.9% poverty			0.95	(0.89,1.02)	1.01	(0.95,1.08)			1.06	(0.99,1.13)
≥20% poverty			1.00	(reference)	1.00	(reference)			1.00	(reference)
CT % college graduates										
40–100			0.71	(0.66,0.76)	0.80	(0.75,0.86)			0.85	(0.79,0.91)
25–39.9			0.82	(0.76,0.87)	0.91	(0.84,0.97)			0.93	(0.87,1.00)
15–24.9			0.87	(0.81,0.92)	0.93	(0.87,0.99)			0.95	(0.89,1.01)
0–14.9			1.00	(reference)	1.00	(reference)			1.00	(reference)

* Parameter estimates for variables in each model adjusted for all other covariates in model (the variables in the column)

** San Francisco (SF) Bay Area = Alameda, Contra Cost, Marin, San Franciso and San Mateo counties

CI = confidence interval

risk factors would have further decreased the magnitude of racial/ethnic disparities in ER status. By the same logic, had we been able to adjust for individual- as well as census tract socioeconomic measures (including across the lifecourse), instead of relying only on the area-based socioeconomic measures, the racial/ethnic disparities in ER status would likely have been further diminished [57, 67–69].

Table 4 Multivariable analysis* of racial/ethnic and socioeconomic disparities in: (a) the odds ratio (OR) versus prevalence ratio (PR) for being ER− versus ER+, based on the imputed data, and (b) the PR for being ER− versus ER+ for the observed versus imputed data, for primary invasive breast cancer among women, San Francisco Bay Area** and Los Angeles County, 1998–2002

Characteristics	Model 1				Model 2			
	Imputed data		Imputed data		Observed data		Imputed data	
	OR	(95% CI)	PR	(95% CI)	PR	(95% CI)	PR	(95% CI)
Race/ethnicity								
White non-Hispanic	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Black non-Hispanic	1.82	(1.65,2.00)	1.47	(1.38,1.56)	1.56	(1.46,1.67)	1.47	(1.38,1.56)
Asian and Pacific Island non-Hispanic	1.11	(1.02,1.21)	1.07	(1.00,1.14)	1.08	(1.00,1.15)	1.07	(1.00,1.14)
Hispanic	1.32	(1.21,1.44)	1.21	(1.14,1.29)	1.25	(1.17,1.34)	1.21	(1.14,1.29)
Age								
0–44	1.76	(1.63,1.90)	1.45	(1.38,1.53)	1.52	(1.43,1.60)	1.45	(1.38,1.53)
45–54	1.45	(1.36,1.55)	1.30	(1.24,1.37)	1.38	(1.31,1.46)	1.30	(1.24,1.37)
≥ 55	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Stage								
Localized	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Regional	0.99	(0.93,1.05)	0.98	(0.93,1.02)	0.92	(0.87,0.96)	0.98	(0.93,1.02)
Remote	0.98	(0.86,1.12)	0.96	(0.89,1.05)	0.93	(0.83,1.04)	0.96	(0.89,1.05)
Tumor size (mm)								
0–10	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
10–20	1.10	(1.00,1.21)	1.08	(1.00,1.17)	1.18	(1.10,1.28)	1.08	(1.00,1.17)
20–30	1.64	(1.48,1.81)	1.48	(1.37,1.59)	1.76	(1.63,1.91)	1.48	(1.37,1.59)
≥30	2.29	(2.07,2.54)	1.83	(1.69,1.97)	2.33	(2.15,2.53)	1.83	(1.69,1.97)
Histology								
Ductal	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Lobular, Type 1	0.26	(0.22,0.32)	0.33	(0.28,0.39)	0.24	(0.20,0.28)	0.33	(0.28,0.39)
Lobular, Type 2	0.29	(0.26,0.34)	0.37	(0.33,0.42)	0.30	(0.27,0.34)	0.37	(0.33,0.42)
Other	1.17	(1.08,1.26)	1.11	(1.06,1.17)	1.07	(1.00,1.13)	1.11	(1.06,1.17)
Catchment area								
SF Bay Area**	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Los Angeles	1.19	(1.12,1.27)	1.12	(1.07,1.17)	1.12	(1.06,1.12)	1.12	(1.07,1.17)
CT poverty/high income measure								
<5% poverty and ≥10% high income	1.16	(1.01,1.32)	1.10	(1.00,1.20)	1.08	(0.98,1.10)	1.10	(1.00,1.20)
<5% poverty and <10% high income	1.19	(1.06,1.34)	1.12	(1.04,1.22)	1.11	(1.01,1.22)	1.12	(1.04,1.22)
5.0–9.9% poverty	1.09	(0.98,1.22)	1.06	(0.98,1.14)	1.05	(0.98,1.14)	1.06	(0.98,1.14)
10–19.9% poverty	1.09	(0.98,1.20)	1.06	(0.99,1.13)	1.06	(0.99,1.14)	1.06	(0.99,1.13)
≥20% poverty	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
CT % college graduates								
40–100	0.79	(0.71,0.87)	0.85	(0.79,0.91)	0.86	(0.79,0.93)	0.85	(0.79,0.91)
25–39.9	0.88	(0.80,0.98)	0.93	(0.87,1.00)	0.96	(0.89,1.04)	0.93	(0.87,1.00)
15–24.9	0.91	(0.83,1.00)	0.95	(0.89,1.01)	0.96	(0.89,1.03)	0.95	(0.89,1.01)
0–14.9	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)

* Parameter estimates for each covariate adjusted for all other covariates in the model (variables in the column)

**San Francisco (SF) Bay Area = Alameda, Contra Cost, Marin, San Franciso and San Mateo counties

CI = confidence interval

The lack of data on health system variables associated with ER status, such as access to and quality of screening and treatment, is also unlikely to have compromised our results, given our inclusion of data on what these health system variables are supposed to affect, e.g., tumor size and stage [1, 33–37]. That said, inclusion of data on health

insurance, delays in obtaining screening, delays in obtaining medical care, and reasons for ER status being unknown, would have been useful for better understanding health system variables affecting ER status. Moreover, racial/ethnic misclassification (likely low for the white non-Hispanic and black non-Hispanic cases [51–55]) is unlikely to have unduly biased the results, since such misclassification is unlikely to have been systematically linked to ER status.

Also meriting caution is our using multiple imputation for the missing data. Justifying our use of this technique, as noted previously, was the inclusion of key known risk factors for ER status, thereby meeting the Missing At Random (MAR) assumption that the probability of missingness depends only on the observed variables [30, 31]. If, however, the data were Not Missing At Random (NMAR, i.e., there are additional *unobserved* predictors of both missingness and ER status), more complex models for non-ignorable non-response are required [30]. Determining whether these assumptions are met depends on conceptual criteria, and cannot be empirically tested in the observed data [29–31].

One additional caveat concerns generalizability, since our study base was restricted to two regions, both within one US state, with cases diagnosed between 1998 and 2002. Our finding thus cannot be generalized to all breast cancer cases in the US for all time periods, especially given secular changes in many of the known risk factors for ER status, including reproductive history, use of hormone therapy, and body mass index[3, 37, 49] and also refinements in assays for ER status [1]. Even so, the results likely do have meaningful implications for the more recent US studies conducted on breast cancer estrogen receptor status and race/ethnicity, e.g., the 21 studies conducted during the past decade [3, 5–15, 20–24].

Interpretation of results

Assuming our results are reasonably valid and reflect the experiences of a reasonably heterogeneous study population, our study raises important questions about the seeming US scientific consensus that intrinsic racial/ethnic disparities exist in breast cancer ER status [1–4]. Our results instead imply this consensus is misleading, since it based predominantly on studies that: (a) lacked socioeconomic data; (b) ignored the problem of missing data; and (c) reported only the odds ratio [3, 6, 7, 9, 12–15, 17, 18, 20, 23], or else a *p*-value for a chi-square test [5, 8, 10, 11, 16]. As with the contrasting prior negative studies, all of which controlled for socioeconomic position and had little or no missing ER data [20–28], we found that taking into account racial/ethnic socioeconomic disparities in ER status and missingness of ER data strongly reduced estimates of racial/ethnic disparities in ER status. Moreover, had we been able to include additional risk factors for ER status known to vary by socioeconomic position within and across racial/ethnic groups, such as hormone therapy, body mass index, and reproductive history [1, 33–35], it is likely that we would have further shrunk the observed racial/ethnic disparities in ER status.

Granted, our study data do not permit us to rule out whether there are particular candidate genes that vary in frequency by race/ethnicity and that shape risk of developing an ER+ versus ER− breast tumor, as some have hypothesized [2–4]. Such a hypothesis, however, would need to account not only for the well-known genetic heterogeneity among the racial/ethnic groups delimited by the official federal US racial/ethnic categories [38, 70–72] but also for why, even within these racial/ethnic groups, socioeconomic disparities exist for risk of being ER+.

Our finding of socioeconomic disparities in ER status even in models containing data on tumor characteristics further implies the existence of additional pathways—other than those captured by tumor size, stage, and histologic type—by which societal conditions influence ER status. Given that ER status remains a powerful predictor of breast cancer survival [1–3, 34, 35], and that research on determinants of ER status remains scant [1, 35, 73, 74], a research program on the social determinants of ER status is warranted. In light of our findings, we emphasize that research on ER status should not be restricted only to cases with known ER status, since doing so would, in the US context, disproportionately include white and more affluent women and exclude women subjected to economic deprivation and women of color. The potential harm to both population health and scientific inference resulting from failing to take into account the full population distribution of exposures and health outcomes and by ignoring socioeconomic confounding has been repeatedly demonstrated, most recently in research on hormone therapy and risk of cardiovascular disease [75, 76], with likely spillover consequences including increased breast cancer incidence attributable to HT use [37, 75–80]. Our findings likewise suggest that better understanding of the determinants of missing ER status and its utility as a health services marker of inadequate medical care [33, 34] would likely be beneficial for efforts to improve breast cancer survival.

A final implication of our study is that research on race/ethnicity and breast cancer estrogen receptor status, like any population health research, requires considering the social as well as biological determinants of health—as well as the social determinants of missingness and data quality. At issue is the conduct not of "politically correct" science, but of *correct* science [38, 77]. Leave out socioeconomic data when studying racial/ethnic health inequities [38, 39, 57, 67–69], or ignore the social patterning of missing data,

[81, 82] and causal inferences are likely to be biased—resulting, in the case of ER status, inflated estimates of racial/ethnic disparities.

Acknowledgments The authors gratefully acknowledge, with permission granted on 20 December 2007, the contributions of Pamela D. Waterman, MPH and Ruihua Yin, MS for their initial work helping to prepare the database used for this study.

References

1. Althuis MD, Fergenbaumm JH, Garcia-Closas M, Brinton LA, Madigan MP, Sherman ME (2004) Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. Cancer Epidemiol Biomarkers Prev 13:1558–1568
2. Ademuyiwa FO, Olopade OI (2003) Racial differences in genetic factors associated with breast cancer. Cancer Metastasis Rev 22:47–53. doi:10.1023/A:1022259901319
3. Chlebowski RT, Chen Z, Anderson GL et al (2005) Ethnicity and breast cancer: factors influencing differences in incidence and outcome. J Natl Cancer Inst 97:439–448
4. Stanford JL, Szklo M, Brinton LA (1986) Estrogen receptors and breast cancer. Epidemiol Rev 8:42–59
5. Morris GJ, Naidu S, Topham AK et al (2007) Differences in breast carcinoma characteristics in newly diagnosed African-American and Caucasian patients—a single-institution compilation compared with the National Cancer Institute's Surveillance, Epidemiology, and End Results Database. Cancer 110:876–884. doi:10.1002/cncr.22836
6. Watlington AT, Byers T, Mouchawar J, Sauaia A, Ellis J (2007) Does having insurance affect differences in clinical presentation between Hispanic and non-Hispanic white women with breast cancer? Cancer 109:2093–2099. doi:10.1002/cncr.22640
7. Woods SE, Luking R, Atkins B, Engel A (2006) Association of race and breast cancer stage. J Natl Med Assoc 98:683–686
8. Cunningham JE, Butler WM (2004) Racial disparities in female breast cancer in South Carolina: clinical evidence for a biological basis. Breast Cancer Res Treat 88:161–176. doi:10.1007/s10549-004-0592-9
9. Porter PL, Lund MJ, Lin MG et al (2004) Racial differences in the expression of cell cycle regulatory proteins in breast carcinoma - Study of young African American and white women in Atlanta, Georgia. Cancer 100:2533–2542. doi:10.1002/cncr.20279
10. Gill KS, Yankaskas BC (2004) Screening mammography performance and cancer detection among black women and white women in community practice. Cancer 100:139–148. doi:10.1002/cncr.11878
11. Ziv E, Tice J, Smith-Bindman R, Shepherd J, Cummings S, Kerlikowske K (2004) Mammographic density and estrogen receptor status of breast cancer. Cancer Epidemiol Biomarker Prev 13:2090–2095
12. Althuis MD, Brogan DD, Coates RJ et al (2003) Breast cancers among very young premenopausal women (United States). Cancer Causes Control 14:151–160. doi:10.1023/A:1023006000760
13. Li CI, Malone KE, Daling JR (2002) Differences in breast cancer hormone receptor status and histology by race and ethnicity among women 50 years of age and older. Cancer Epidemiol Biomarker Prev 11:601–607
14. Miller BA, Hankey BF, Thomas TL (2002) Impact of sociodemographic factors, hormone receptor status, and tumor grade on ethnic differences in tumor stage and size for breast cancer in US women. Am J Epidemiol 155:534–545. doi:10.1093/aje/155.6.534
15. Furberg H, Millikan R, Dressler L, Newman B, Geradts J (2001) Tumor characteristics in African American and white women. Breast Cancer Res Treat 68:33–43. doi:10.1023/A:1017994726207
16. Elledge RM, Clark GM, Chamness GC, Osborne CK (1994) Tumor biologic factors and breast-cancer prognosis among White, Hispanic, and Black women in the United-States. J Natl Cancer Inst 86:705–712. doi:10.1093/jnci/86.9.705
17. Harlan LC, Coates RJ, Block G et al (1993) Estrogen receptor status and dietary intakes in breast cancer patients. Epidemiology 4:25–31. doi:10.1097/00001648-199301000-00006
18. Stanford JL, Greenberg RS (1989) Breast cancer incidence in young women by estrogen receptor status and race. Am J Public Health 79:71–73
19. Stanford JL, Szklo M, Boring CC et al (1987) A case-control study of breast cancer stratified by estrogen receptor status. Am J Epidemiol 125:184–194
20. Martinez ME, Nielson CM, Nagle R, Lopez AM, Kim C, Thompson P (2007) Breast cancer among Hispanic and non-Hispanic white women in Arizona. J Health Care Poor Underserved 18:130–145. doi:10.1353/hpu.2007.0112
21. Menes TS, Ozao J, Kim U (2007) Breast cancer and ethnicity: strong association between reproductive risk factors and estrogen receptor status in Asian patients—a retrospective study. Breast J 13:352–358. doi:10.1111/j.1524-4741.2007.00442.x
22. Gordon N (2003) Socioeconomic factors and breast cancer in black and white Americans. Cancer Metastasis Rev 22:55–65. doi:10.1023/A:1022212018158
23. Elmore JG, Moceri VM, Carter D, Larson EB (1998) Breast carcinoma tumor characteristics in black and white women. Cancer 83:2509–2515. doi:10.1002/(SICI)1097-0142(19981215)83:12<2509::AID-CNCR15>3.0.CO;2-V
24. Krieger N, van den Eeden SK, Zava D, Okamoto A (1997) Race/ethnicity, social class, and prevalence of breast cancer prognostic biomarkers: a study of white, black, and Asian women in the San Francisco Bay Area. Ethn Dis 7:137–149
25. Weiss SE, Tartter PI, Ahmed S et al (1995) Ethnic differences in risk and prognostic factors for breast cancer. Cancer 76:268–274. doi: 10.1002/1097-0142(19950715)76:2<268::AID-CNCR2820760217>3.0.CO;2-1
26. Gordon NH (1995) Association of education and income with estrogen receptor status in primary breast cancer. Am J Epidemiol 142:796–803
27. Chen VW, Correa P, Kurman RJ et al (1994) Histological characteristics of breast carcinoma in blacks and whites. Cancer Epidemiol Biomarkers Prev 3:127–135
28. Ansell D, Whitman S, Lipton R, Cooper R (1993) Race, income, and survival from breast cancer at two public hospitals. Cancer 72:2974–2978. doi:10.1002/1097-0142(19931115)72:10<2974::AID-CNCR2820721019>3.0.CO;2-M
29. Crowe JP Jr, Gordon NH, Hubay CA, Pearson OH, Marshall JS, McGuire WL (1986) The interaction of estrogen receptor status and race in predicting prognosis for stage I breast cancer patients. Surgery 100:599–605
30. Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall, Boca Raton
31. Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142:1255–1264
32. Rubin D (1976) Inference and missing data. Biometrika 63:581–592. doi:10.1093/biomet/63.3.581
33. Haggstrom DA, Quale C, Smith-Bindman R (2005) Differences in the quality of breast cancer care among vulnerable populations. Cancer 104:2347–2358. doi:10.1002/cncr.21443
34. Cross CK, Harris J, Recht A (2002) Race, socioeconomic status, and breast carcinoma in the U.S: what have we learned from clinical studies. Cancer 95:1988–1999. doi:10.1002/cncr.10830

35. Hwang ES, Chew T, Shiboski S, Farren G, Benz CC, Wrensch M (2005) Risk factors for estrogen receptor-positive breast cancer. Arch Surg 140:58–62. doi:10.1001/archsurg.140.1.58
36. Thomson CS, Hole DJ, Twelves CJ, Brewster DH, Black RJ, Scottish Cancer Therapy Network (2001) Prognostic factors in women with breast cancer: distribution by socioeconomic status and effect on differences in survival. J Epidemiol Community Health 55:308–315. doi:10.1136/jech.55.5.308
37. Krieger N (2008) Hormone therapy and the rise and perhaps fall of US breast cancer incidence rates: critical reflections. Int J Epidemiol 37:627–637. doi:10.1093/ije/dyn055
38. Krieger N (2005) Stormy weather: "race", gene expression, and the science of health disparities. Am J Public Health 95:2155–2160. doi:10.2105/AJPH.2005.067108
39. Krieger N (2003) Does racism harm health? did child abuse exist before 1962?—on explicit questions, critical science, and current controversies: an ecosocial perspective. Am J Public Health 93:194–199
40. Ernst W, Harris B (eds) (1999) Race, science, and medicine, 1700–1960. Routledge, London
41. Krieger N (1994) Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med 39:887–903. doi:10.1016/0277-9536(94)90202-X
42. Krieger N (2004) Ecosocial theory. In: Anderson N (ed) Encyclopedia of health and behavior. Sage, Thousand Oaks, CA, pp 292–294
43. Krieger N (2008) What's level got to do with it?—proximal, distal, and the politics of causation. Am J Public Health 98:221–230. doi:10.2105/AJPH.2007.111278
44. Davies HT, Crombie IK, Tavakoli M (1998) When can odds ratios mislead? BMJ 316:989–991
45. Zocchetti C, Consonni D, Bertazzi PA (1997) Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. Int J Epidemiol 26:220–223. doi:10.1093/ije/26.1.220
46. Greenland S, Thomas DC, Morgenstern H (1986) The rare-disease assumption revisited: a critique of "estimators of relative risk for case-control studies". Am J Epidemiol 124:869–883
47. Northern California Cancer Center (2008) http://www.nccc.org/site/c.foJNIXOyEpH/b.2577075/k.BE87/Home.htm. Accessed 22 May 2008
48. Los Angeles Cancer Surveillance Program (2008) http://www.usc.edu/schools/medicine/departments/preventive_medicine/divisions/epidemiology/research/csp/index.html. Accessed 22 May 2008
49. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Yin R, Coull BA (2006) Race/ethnicity and changing US socioeconomic gradients in breast cancer incidence: California and Massachusetts, 1978–2002. Cancer Causes Control 17:217–226. doi:10.1007/s10552-005-0408-1
50. US Office of Management and Budget (2008) Revisions to the standards for the classification of federal data on race and ethnicity. Federal register notice, October 30, 1997. http://www.whitehouse.gov/omb/fedreg/1997standards.html. Accessed 22 May 2008
51. Stewart SL, Swallen KC, Glaser SL, Horn-Ross PL, West DW (1998) Adjustment of cancer incidence rates for ethnic misclassification. Biometrics 54:774–781. doi:10.2307/3109783
52. Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S (2005) Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. Ann Epidemiol 15:71–79. doi:10.1016/j.annepidem.2004.03.002
53. Stewart SL, Swallen KC, Glaser SL, Horn-Ross PL, West DW (1999) Comparison of methods for classifying Hispanic ethnicity in a population-based cancer registry. Am J Epidemiol 149:1063–1071
54. Swallen KC, West DW, Stewart SL, Glaser SL, Horn Ross PL (1997) Predictors of misclassification of Hispanic ethnicity in a population-based cancer registry. Ann Epidemiol 7:200–206. doi:10.1016/S1047-2797(96)00154-8
55. Clegg LX, Reichman ME, Hankey BF 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 18:177–187
56. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW (2001) On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 91:1114–1116
57. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV (2005) Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the *Public Health Disparities Geocoding Project*. Am J Public Health 95:312–323. doi:10.2105/AJPH.2003.032482
58. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV (2003) Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—*The Public Health Disparities Geocoding Project*. Am J Public Health 93:1655–1671
59. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R (2002) Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: *The Public Health Disparities Geocoding Project*. Am J Epidemiol 156:471–482. doi:10.1093/aje/kwf068
60. US Bureau of the Census. Poverty areas (2008) http://www.census.gov/population/socdemo/statbriefs/povarea.html. Accessed 22 May 2008.61
61. King G, Honaker J, Joseph A, Scheve K (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am Pol Sci Review 95: 49–69
62. Spiegelman D, Hertzmark E (2005) Easy SAS calculations for risk or prevalence ratios or differences. Am J Epidemiol 162:199–200. doi:10.1093/aje/kwi188
63. Greenland S (2004) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am J Epidemiol 160:301–305. doi:10.1093/aje/kwh221
64. McNutt LA, Wu C, Xue X, Hafner JP (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol 157:940–943. doi:10.1093/aje/kwg074
65. Skov T, Deddens J, Petersen MR, Endahl L (1998) Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol 27:91–95. doi:10.1093/ije/27.1.91
66. SAS Institute (2001) SAS language reference, Version 8. SAS Institute, Cary, NC
67. Krieger N, Williams D, Moss N (1997) Measuring social class in US public health research: concepts, methodologies and guidelines. Annu Rev Public Health 18:341–378. doi:10.1146/annurev.publhealth.18.1.341
68. Davey Smith G (2000) Learning to live with complexity: ethnicity, socioeconomic position, and health in Britain and the United States. Am J Public Health 90:1694–1698
69. Galobardes B, Lynch J, Davey Smith G (2007) Measuring socioeconomic position in health research. Br Med J 81–82:21–37
70. Goodman AH (2000) Why genes don't count (for racial differences in health). Am J Public Health 90:1699–1702
71. Parra EJ, Kittles RA, Shriver MD (2004) Implications of correlations between skin color and genetic ancestry for biomedical research. Nature Genetics 36(suppl):S54–S60. doi:10.1038/ng1440
72. Feldman MW, Lewontin RC, King MC (2003) Race: a genetic melting-pot. Nature 424:374. doi:10.1038/424374a
73. Fowler AM, Alarid ET (2007) Amping up estrogens receptors in breast cancer. Breast Cancer Res 9:305 (doi:10.1186/box1748). doi:10.1186/bcr1748

74. Zhu K, Bernard LJ, Levine RS, Williams SM (1997) Estrogen receptor status of breast cancer: a marker of different stages of tumor or different entities of the disease? Med Hypotheses 49:69–75. doi:10.1016/S0306-9877(97)90255-3
75. Petitti D (2004) Commentary: hormone replacement therapy and coronary heart disease: four lessons. Int J Epidemiol 33:461–463. doi: 10.1093/ije/dyh192
76. Lawlor DA, Davey Smith G, Ebrahim S (2004) Commentary: The hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol 33:464–467. doi:10.1093/ije/dyh124
77. Krieger N (2007) Why epidemiologists cannot afford to ignore poverty. Epidemiology 18:658–663
78. Prentice RL, Chlebowski RT, Stefanick ML, Mansons JE, Langer RD, Pettinger M, Hendrix SL, Hubbell FA, Kooperberg C, Kuller LH, Lane DS, McTiernan A, O'Sullivan MJ, Rossouw JE, Anderson GL (2008). Conjugated equine estrogens and breast cancer risk in the Women's Health Initiative clinical trial and observational study. Am J Epidemiol 167:1407–1415. doi:10.1093/aje/kwn090
79. Clarke CA, Glaser SL (2007) Declines in breast cancer after the WHI: apparent impact of hormone therapy. Cancer Causes Control 18:847–852. doi:10.1007/s10552-007-9029-1
80. Ravdin PM, Cronin KA, Howlader N, Berg CD, Chlebowski RT, Feuer EJ, Edwards BK, Berry DA (2007) The decrease in breast-cancer incidence in 2003 in the United States. New Engl J Med 356:1670–1674. doi:10.1056/NEJMsr070105
81. Kim S, Egerter S, Cubbin C, Takahashi ER, Braveman P (2007) Potential implications of missing income data in population-based surveys: an example from a postpartum survey in California. Public Health Reports 122:753–764
82. Chen JT, Kaddour A, Krieger N (2008) Re: Kim et al, "Potential Implications of missing income data in population-based surveys". Public Health Reports 123:260