Overcoming the Absence of Socioeconomic Data in Medical Records: Validation and Application of a Census-Based Methodology

Nancy Krieger, PhD

ABSTRACT

Background. Most US medical records lack socioeconomic data, hindering studies of social gradients in health and ascertainment of whether study samples are representative of the general population. This study assessed the validity of a census-based approach in addressing these problems.

Methods. Socioeconomic data from 1980 census tracts and block groups were matched to the 1985 membership records of a large prepaid health plan (n = 1.9 million), with the link provided by each individual's residential address. Among a subset of 14 420 Black and White members, comparisons were made of the association of individual, census tract, and census block-group socioeconomic measures with hypertension, height, smoking, and reproductive history.

Results. Census-level and individual-level socioeconomic measures were similarly associated with the selected health outcomes. Census data permitted assessing response bias due to missing individual-level socioeconomic data and also contextual effects involving the interaction of individual- and neighborhood-level socioeconomic traits. On the basis of block-group characteristics, health plan members generally were representative of the total population; persons in impoverished neighborhoods, however, were underrepresented.

Conclusions. This census-based methodology offers a valid and useful approach to overcoming the absence of socioeconomic data in most US medical records. (*Am J Public Health.* 1992;92:703–710)

Introduction

One problem frequently encountered by public health researchers and health planners in the United States is the absence of socioeconomic data in many widely used and routinely collected sources of health and disease information, such as medical records, disease registries, and vital statistics.[1,2] This omission is important, because social class, education, and income are known to influence both health status and use of health services.[1–7] Without these data, it is difficult not only to assess the role of socioeconomic factors in shaping the public's health, but also to ascertain whether a given study population is representative of the general population, and hence whether the study's findings legitimately can be generalized to the public at large.

To overcome this obstacle, various US investigators have sought to supplement individual-level health records with social class measures derived from an additional and easily available source of socioeconomic data, the US census.[1,8–20] In this approach, individuals are characterized by the socioeconomic profile of the immediate neighborhood in which they live (usually defined as the census tract), with the link provided by each person's residential address. To date, however, little research has examined the validity of this methodology,[1,8,21] including the degree to which it may result in ecologic fallacy (i.e., the erroneous inference of causal relations at the individual level on the basis of grouped data).[8,22–23]

To evaluate this census-based approach, I obtained health records with individual-level socioeconomic data from a large, multiethnic prepaid health plan and linked each individual's record to census-derived socioeconomic data acquired from two different geographic levels: (1) the census tract and (2) the census block group (a smaller and often more homogenous subdivision of the census tract that, on average, contains 1000 persons, as compared with the 4000 that typically reside in a census tract[24]). I then compared the association of both individual-level and census-based socioeconomic measures with four different health characteristics known to vary by race and socioeconomic position: hypertension,[2,25] height,[2,7,26] cigarette smoking,[2,7] and number of full-term pregnancies.[2,27] Census-based socioeconomic measures also were used to assess whether members of this health plan are representative of the surrounding general public. These analyses extend a smaller pilot project that compared individual, household, and census-derived social class measures as correlates of Black–White differences in reproductive history.[21]

Methods

Population

Study subjects consisted of all persons contained in the June 1985 computer-stored membership tape of the Kaiser Permanente Medical Care Program (KPMCP), Northern California Region (n = 1 924 995). Only data on age and gender are available for the total KPMCP membership; race/ethnicity and socioeco-

The author is with the Division of Research, Kaiser Foundation Research Institute, Oakland, Calif.

Requests for reprints should be sent to Nancy Krieger, PhD, Division of Research, Kaiser Foundation Research Institute, 3451 Piedmont Avenue, Oakland, CA 94611.

This paper was submitted to the Journal April 2, 1991, and accepted with revisions August 22, 1991.

nomic position are not ascertained. In 1987, the Medical Economics Department of KPMCP sent the June 1985 membership tape to be "geocoded" by a commercial firm, at a cost of $4.50 per 1000 records. This process assigned each individual a county code, census tract number, and block-group number based on the member's residential address and on geographic regions defined in the 1980 census.

A total of 1 593 388 KPMCP members (82.8%) were successfully geocoded. These members lived within 22 counties that were home to 98.8% of the total 1985 Northern California Region KPMCP membership. Missing or incorrect addresses, along with the absence of block-group codes in the 1980 census for three small Northern California counties (Amador, Lake, and Nevada), precluded fully geocoding the remainder. Members who were and were not successfully geocoded did not differ by gender and only differed slightly by age.

Census-Derived Data

Census tract and census block-group data from the 1980 US census were matched to each geocoded member's individual record. Each tract and block group was characterized by its social class and race/ethnic composition, and also by poverty and educational level.

Predominantly working-class tracts and block groups were defined as neighborhoods where 66% or more of the employed population belonged to the following census-based occupational categories: clerical and administrative support, sales, private household and other service occupations (except protective services), craft, transportation, and laborers. These occupations were selected because they disproportionately contain people who can be considered working class (i.e., employees who do not own their workplace, are not self-employed, and generally occupy subordinate positions at work).[21,28] The remaining census-defined occupations were considered non–working class: executive, administrative, and managerial; professional specialty; technicians and related support; protective service; and farming, forestry, and fishing (which includes farm owners and managers).

Impoverished tracts and block groups consisted of federally defined poverty areas, in which 20% or more of the population lived below the poverty line.[29] This economic indicator was chosen over the more commonly used, statistically defined measure of median family income because it takes into account a family's size and age structure and also pertains to the ability to buy a specified market basket of goods.[1,30,31] In 1980, the poverty level was set at $7356 for a family of two adults and two children.[29] Undereducated tracts and block groups were defined as regions where at least 25% of persons 25 years old or more had not completed high school.

Multiphasic Health Check-up Examination Data

Data from the multiphasic health check-up (MHC) examination were available for 17 200 persons contained in the geocoded June 1985 membership tape. This group represented 91.0% of the 18 904 KPMCP members who completed both the questionnaire and laboratory portions of the MHC examination in 1985 and who were KPMCP members as of June 1985.

Data obtained from the MHC examination included age at MHC examination, gender, race, years of completed education, occupation (current, if employed, or last occupation, if unemployed or retired), diastolic and systolic blood pressure, height, weight, cigarette-smoking history, and, for women, number of full-term pregnancies (NFTP). No data on income were collected. Using the same occupational groupings described above, social class was categorized as working class or non–working class. High blood pressure (HBP) was defined as diastolic ($\geq$ 90 mm Hg) and/or systolic ($\geq$ 145 mm Hg).[32] The Quetelet index (weight/height^2) served as a measure of body mass. The final sample consisted of 8674 White and 5566 Black KPMCP members included among the 17 055 persons in the MHC subset for whom race was identified.

Statistical Methods

Comparisons of demographic and health-related characteristics used the t test for continuous variables and the χ^2 test for categorical variables. The association of socioeconomic measures with the four health characteristics was determined by multivariate regression models run separately for the individual, tract, and block-group data. Analyses were performed with SAS programs for the personal computer.[33]

Multiple linear regression was used for health-related outcomes treated as continuous variables (height, NFTP). These models, in the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

permit estimations of how much the dependent variable (y) increases or decreases with a unit change in each independent variable (x_k), as indicated by its coefficient β_k, controlling for the other independent variables in the model.[33,34] Logistic regression was used for dichotomous (yes or no) outcomes (HBP, smoking status). This method compares the probability (p) of experiencing an outcome given a particular exposure x_k, and is expressed in the form:[33,35]

$$\text{logit}(p) = \log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

The relative risk associated with a unit change in each independent variable x_k, again controlling for other independent variables included in the model, can be calculated as $e^{\beta k}$.

For each health characteristic, the same group of individuals was analyzed at the individual, tract, and block-group level, with all results adjusted for the same relevant individual-level risk factors (e.g., age and gender). These analyses are referred to as the complete models, since they included only those individuals missing no individual-level or census-derived data for any of the variables included at any of the three model levels. Total models permitted analyzing the effect of response bias due to missing individual-level socioeconomic data, and included persons missing these data but not census-level socioeconomic data or other relevant individual-level data. Contextual regression analyses,[36,37] which incorporate individual- and group-level measures of the same traits (e.g., individual class position and working-class composition of that individual's tract or block group), were used to determine whether the effects of individual socioeconomic position are modified by neighborhood conditions. These models used the individual- and census-level data in the form of a combined dummy variable.

On the basis of the results of these analyses, block-group characteristics of the total geocoded KPMCP membership were compared with those of the total population for all 22 counties combined. Because of the extremely large size of the KPMCP sample (1.6 million, representing 21.6% of the total population), statistical tests were not used. Instead, differences between these two groups are noted and assessed.

TABLE 1—Comparison of Demographic and Health-Related Characteristics

	White		Black		Comparison of Blacks to Whites[b] (95% Confidence Interval)
	Value[a]	% missing data	Value[a]	% missing data	
Age, y (mean ± SD)	51.8 ± 16.3	0.0	47.8 ± 15.4	0.0	−4.0 (−5.3, −2.7)
Gender: women, %	54.2	0.0	60.3	0.0	1.11 (1.07, 1.14)
Class: working class, %	45.2	41.0	63.7	39.8	1.41 (1.36, 1.47)
Education: less than high school, %	22.7	24.9	38.6	25.6	1.70 (1.60, 1.80)
Height, cm (mean ± SD)	168.8 ± 9.9	0.1	168.1 ± 9.4	0.1	−0.7 (−1.0, −0.4)
Weight, kg (mean ± SD)	72.4 ± 15.4	0.1	78.1 ± 17.2	0.0	5.7 (5.1, 6.3)
Blood pressure, mm Hg					
Diastolic (mean ± SD)	75.3 ± 9.6	0.2	78.6 ± 10.4	0.2	3.3 (3.0, 3.6)
Systolic (mean ± SD)	130.7 ± 19.4	0.1	133.3 ± 19.1	0.1	2.6 (2.0, 3.3)
High Blood pressure[c]: yes, %	28.3	0.0	35.1	0.0	1.24 (1.18, 1.30)
Ever smoke: yes, %	51.7	24.9	53.1	25.2	1.03 (0.94, 1.07)
Still smokes: yes, %	33.8	4.1[d]	53.4	3.1[d]	1.58 (1.48, 1.68)
Full-term pregnancies (FTP) (mean ± SD)	1.6 ± 1.5	25.2[e]	2.1 ± 2.0	24.9[e]	0.5 (0.4, 0.6)
Census tract					
66+% working class, %	35.1	2.1	77.4	0.8	2.21 (2.14, 2.28)
25+% less than high school, %	20.4	2.8	67.2	0.8	3.29 (3.14, 3.44)
20+% less than poverty, %	10.2	2.1	43.0	0.8	4.22 (3.94, 4.53)
Census block group					
66+% working class, %	36.7	0.2	76.9	0.1	2.10 (2.03, 2.16)
25+% less than high school, %	24.5	0.2	66.6	0.1	2.72 (2.61, 2.83)
20+% less than poverty, %	11.1	0.2	43.3	0.1	3.90 (3.65, 4.17)

Note. 8674 White and 5566 Black KPMCP members.
[a]Value among those not missing data for the specified variable.
[b]Comparison: absolute difference (Black–White) for continuous variables, relative prevalence (Black ÷ White) for categorical variables.
[c]High blood pressure defined as diastolic (≥90 mm Hg) and/or systolic (≥145 mm Hg).
[d]Universe: all ever smokers (3371 Whites and 2210 Blacks).
[e]Universe: all women (4700 Whites and 3358 Blacks).

Results

Characteristics of the MHC Subset

The extent of missing data varied considerably for the demographic and health-related traits recorded at the MHC examination (Table 1). With the exception of the few persons missing tract-level data, Black and White MHC study subjects did not significantly differ in the proportion missing data for any given trait. They did differ significantly, however, for every demographic and health-related outcome except one (percentage ever smoked) (Table 1).

Validation of Census-Based Socioeconomic Measures

Among persons included in the complete models, measures of social class and educational attainment at the individual, census-tract, and block-group levels yielded similar estimates for the association of these socioeconomic variables with, respectively, HBP, height, still smoking (among ever smokers), and NFTP (Table 2). Block-group models generally yielded tighter confidence intervals than tract models. Tract and block-group models both tended to underestimate slightly the socioeconomic associations observed at the individual level, and neither rejected the null hypothesis (no class effects) when this hypothesis was not rejected by individual-level models. For each outcome, Black–White differences were similarly reduced by adjusting for individual- and census-level socioeconomic measures (Table 2).

Use of Census-Based Measures to Assess Respondent Bias

Adding persons who lacked individual- but not census-level socioeconomic data markedly increased the number of available subjects (Table 3). These additional persons tended to be older, and were more likely to be women and to reside in predominantly working-class, impoverished, and undereducated block groups, than persons included only in the complete models.

Evidence of respondent bias in the complete models was most apparent for Black–White comparisons concerning HBP and NFTP (Table 3). The adjusted relative risk of HBP among Black as compared with White MHC study subjects was 12% lower in the total than in the complete models (1.5 vs 1.7), while the relative excess of NFTP was 20% lower (0.4 vs 0.5). As would be expected, including substantially more persons in the total models produced tighter confidence intervals.

Contextual Analyses

Contextual regression analyses were conducted only with persons included in the complete models, since only these subjects possessed both individual- and census-level social class or educational measures. Suggestive contextual effects for both measures were observed for all four outcomes in both tract and block-group models. For example, as shown in Table 4, working-class women who lived in non–working-class block groups had virtually the same number of full-term pregnancies as non–working-class women who likewise lived in non–working-class block groups, whereas working-class women who lived in predominantly working-class block groups had 0.4 more full-term pregnancies than this latter group ($P < .05$).

Application of the Census-Based Methodology

A comparison of block-group characteristics of the geocoded 1985 KPMCP membership and the 1980 general population in the 22 corresponding counties re-

TABLE 2—Association of Race with Dependent Variables

	Model Characteristics					Risk Estimates[a] (95% Confidence Intervals)			
	Type of Regression Model	N White	N Black	Models Adjusted for	Model level	Race	Race +	Class +	Education
HBP[b]	Logistic	4960	3256	Age	Individual	1.9 (1.7, 2.1)	1.8 (1.6, 2.0)	1.0 (0.9, 1.2)	1.3 (1.2, 1.5)
				Gender	Tract	1.9 (1.7, 2.1)	1.7 (1.5, 1.9)	1.0 (0.9, 1.2)	1.2 (1.0, 1.4)
				Quetelet index	Block group	1.9 (1.7, 2.1)	1.7 (1.5, 1.9)	1.1 (0.9, 1.3)	1.2 (1.0, 1.4)
Height, cm	Linear	4975	3266	Age	Individual	−0.4 (−0.7, 0.1)	−0.1 (−0.4, 0.2)	−0.7 (−1.0, −0.4)	−1.1 (−1.4, −0.7)
				Gender	Tract	−0.4 (−0.7, −0.1)	−0.0 (−0.3, 0.4)	−0.4 (−0.9, −0.0)	−0.5 (−0.9, −0.0)
					Block group	−0.4 (−0.7, −0.1)	0.0 (−0.3, 0.4)	−0.7 (−1.0, 0.3)	−0.4 (−0.8, 0.0)
Smokes	Logistic	2544	1703	Age	Individual	2.1 (1.8, 2.4)	1.9 (1.7, 2.2)	1.4 (1.2, 1.6)	1.4 (1.2, 1.7)
				Gender	Tract	2.1 (1.8, 2.4)	1.8 (1.5, 2.1)	1.2 (1.0, 1.4)	1.2 (1.0, 1.5)
					Block group	2.1 (1.8, 2.4)	1.8 (1.6, 2.1)	1.2 (1.0, 1.4)	1.2 (1.0, 1.4)
NFTP	Linear	2535	1932	Age	Individual	0.7 (0.6, 0.8)	0.6 (0.5, 0.7)	0.1 (0.0, 0.2)	0.4 (0.3, 0.6)
					Tract	0.7 (0.6, 0.8)	0.5 (0.4, 0.7)	0.2 (0.0, 0.4)	0.1 (−0.1, 0.3)
					Block group	0.7 (0.6, 0.8)	0.5 (0.4, 0.7)	0.2 (0.0, 0.3)	0.2 (0.0, 0.3)

Note. HBP = high blood pressure; NFTP = number of full-term pregnancies.
[a]Comparison groups:
Race: Black vs White
Class: Individual = working class vs non-working class
Census tract = 66+% working-class vs <66% working-class census tract
Block group = 66+% working-class vs <66% working-class census block group
Education: Individual = <high school vs. ≥high school
Census tract = 25+% <high school vs <25% <high school census tract
Block group = 25+% <high school vs <25% <high school census block group
For linear regression: estimate of unit change in outcome associated with unit change in race, class, and education measures
For logistic regression: estimate of relative risk of outcome associated with unit change in race, class, and education measures
[b]HBP defined as diastolic (≥90 mm Hg) and/or systolic (≥ 145 mm Hg).

vealed few differences with regard to race/ethnic composition (Table 5). Although a comparable proportion of KPMCP members and the general population lived in predominantly working-class block groups, KPMCP members were somewhat less likely than the general public to live in impoverished and undereducated block groups (Table 6).

Discussion

The census-based methodology of supplementing individuals' health records with socioeconomic data from their census block groups appears to offer a valid and useful approach to overcoming the absence of socioeconomic data in US medical records. Among 14 240 Black and White men and women who took the KPMCP MHC examination, individual, tract, and block-group measures of social class and education provided highly comparable estimates of association with four diverse health characteristics known to exhibit marked social class and race/ethnic gradients[2,7,25–27]: hypertension, height, smoking, and number of full-term pregnancies.

Beyond this, the use of census data permitted assessing the effect of respondent bias introduced by missing individual-level data, and also allowed contextual analyses to be conducted. These analyses indicated that census data may augment, and not simply approximate, individual-level socioeconomic data. Lastly, the census-based methodology provided a straightforward and relatively inexpensive way of ascertaining whether the KPMCP membership is representative of the surrounding general population, with the findings recently corroborated by the 1989 to 1991 California Risk Factor Surveillance Study (N. Gordon, written communication, February 1991).

Comparison with Other Studies Using Census-Derived Data

To date, very few US studies have compared individual-level and census-derived socioeconomic indicators as correlates of health status.[1,8,21] The greater heterogeneity among residents of census tracts as compared with block groups, however, has long been of concern,[38,39] with some evidence suggesting that block-group data can identify pockets of poverty or affluence not apparent at the tract level.[40] Also bolstering this study's findings are the results of a related pilot project carried out in Alameda County, California.[21] This study of 51 Black and 50 White women found that census block-group measures of social class and poverty closely approximated individual-level measures as correlates of women's reproductive histories, whereas comparable data from the tract level performed less well. Contextual analyses likewise indicated the importance of categorizing women by both individual-level and block-group–level socioeconomic characteristics.[21]

Additional evidence regarding contextual effects and health status stems from another Alameda County study that found that persons who lived in poverty areas were 1.7 times more likely to die during 9 years of follow-up than were persons living in nonpoverty areas, even after controlling for a large number of demographic, physiological, and psychosocial variables.[41] Other US research has observed contextual effects for such diverse phenomena as voting behavior,[42–43] students' test scores,[44] adult participation in community-based organizations,[45] and social problems associated with residential segregation[46] (e.g., Black professionals in the United States are more likely to live in working-class neighborhoods than their White peers, such that analyses "controlling" for only individual-level class differences fail to capture important neighborhood-related differences in socioeconomic position).

TABLE 3—Comparison of Complete and Total Models

Dependent Variable	Model characteristics: Type of Regression Model	Models Adjusted for	Block-Group Model Level	Number in Model: Total	White	Black	Risk Estimates[a] (95% Confidence Intervals): Race	Race + Class + Education
HBP[b]	Logistic	Age Gender Quetelet index	Complete	8216	4960	3256	1.9 (1.7, 2.1)	1.7 (1.5, 1.9) / 1.1 (0.9, 1.3) / 1.2 (1.0, 1.4)
			Total	14 159	8623	5536	1.7 (1.5, 1.8)	1.5 (1.4, 1.7) / 1.0 (0.9, 1.2) / 1.2 (1.0, 1.4)
			Increase, %[c]	72.3	73.9	70.0		
Height, cm	Linear	Age Gender	Complete	8241	4975	3266	−0.4 (−0.7, −0.1)	0.0 (−0.3, 0.4) / −0.7 (−1.0, −0.3) / −0.4 (−0.8, 0.0)
			Total	14 198	8645	5553	−0.4 (−0.6, −0.2)	0.0 (−0.2, 0.2) / −0.6 (−0.9, −0.3) / −0.5 (−0.8, −0.2)
			Increase, %[c]	72.3	73.8	70.0		
Smokes	Logistic	Age Gender	Complete	4247	2554	1703	2.1 (1.8, 2.4)	1.8 (1.6, 2.1) / 1.2 (1.0, 1.4) / 1.2 (1.0, 1.4)
			Total	5365	3228	2137	2.1 (1.8, 2.3)	1.8 (1.6, 2.0) / 1.2 (1.1, 1.4) / 1.2 (1.0, 1.4)
			Increase, %[c]	26.3	26.4	25.5		
NFTP	Linear	Age	Complete	4467	2535	1932	0.7 (0.6, 0.8)	0.5 (0.4, 0.7) / 0.2 (0.0, 0.3) / 0.2 (0.0, 0.3)
			Total	6035	3514	2521	0.6 (0.5, 0.7)	0.4 (0.3, 0.6) / 0.2 (0.0, 0.3) / 0.2 (0.0, 0.3)
			Increase, %[c]	35.1	38.6	30.5	30.5	

Note. HBP = high blood pressure; NFTP = number of full-term pregnancies.
[a]Comparison groups:
Race: Black vs White
Class: 66 + % working class vs <66% working class Black group
Education: 25 + % < high school vs. <25% < high school Black group
For linear regression: estimate of unit change in outcome associated with unit change in race, class, and education measures
For logistic regression: estimate of relative risk of outcome associated with unit change in race, class, and education measures
[b]HBP defined as diastolic (≥90 mm Hg) and/or systolic (≥ 145 mm Hg).
[c]Percentage increase in number of persons in total as compared with complete model.

TABLE 4—Association of Race and Contextual Social Class Measures with Dependent Variables

Dependent Variable	Model Characteristics: Type of Regression Model	Number in Model: White	Black	Model Adjusted for	Risk Estimates[a] (95% Confidence Intervals): Race	NWC(I) + <66% WC(BG) vs NWC(I) + 66+% WC (BG)	NWC(I) + <66% WC(BG) vs WC(I) + <66% WC (BG)	NWC(I) + <66% WC(BG) vs WC(I) + 66+% WC (BG)
HBP[b]	Logistic	4960	3256	Age Gender Quetelet index	1.7 (1.5, 1.9)	1.2 (1.0, 1.4)	1.1 (1.0, 1.3)	1.3 (1.1, 1.5)
Height, cm	Linear	4975	3266	Age Gender	0.0 (−0.3, 0.4)	−0.9 (−1.3, −0.4)	−1.0 (−1.4, −0.6)	−1.6 (−2.0, −1.2)
Smokes	Logistic	2544	1703	Age Gender	1.8 (1.6, 2.1)	1.4 (1.1, 1.7)	1.6 (1.3, 2.0)	1.8 (1.5, 2.2)
NFTP	Linear	2535	1932	Age	0.6 (0.4, 0.7)	0.2 (0.0, 0.3)	0.1 (−0.1, 0.3)	0.4 (0.2, 0.6)

Note. NWC(I) = non–working class (individual); <66% WC(BG) = less than 66% working class (block group); 66+% WC(BG) = 66% or more working class (block group); WC(I) = working class (individual); HBP = high blood pressure;NFTP = number of full-term pregnancies.
[a]Comparison groups:
Race: Black vs White
For linear regression: estimate of unit change in outcome associated with unit change in race and contextual class measures
For logistic regression: estimate of relative risk of outcome associated with unit change in race and contextual class measures
[b]HBP defined as diastolic (≥90 mm Hg) and/or systolic (≥ 145 mm Hg).

TABLE 5—Demographic Comparison of Block-Group Race/Ethnic Composition of 1985 Kaiser Permanente (KP) Membership and 1980 General Population (GP): 22 Selected Counties in Northern California

Block-Group Composition	Percentage Living in Block Groups with Specified Race/Ethnic Composition														
	White			Black			Native American			Asian/Pacific Islander			Hispanic Origin[a]		
	KP	GP	KP/GP	KP	GP	KP/GP	KP	GP	KP/GP	KP	GP	KP/GP	KP	GP	KP/GP
<20%	4.1	3.7	1.11	88.5	90.7	0.98	100.0	100.0	1.00	89.7	91.4	0.98	84.1	81.4	1.03
20%–39%	4.8	4.4	1.09	4.5	3.7	1.22	0.0	0.0	...	8.1	6.6	1.23	11.5	12.4	0.93
40%–59%	10.0	9.2	1.09	2.4	2.0	1.20	0.0	0.0	...	1.9	1.6	1.19	3.4	4.4	0.77
60%–79%	23.1	21.4	1.08	2.1	1.7	1.24	0.0	0.0	...	0.2	0.2	1.00	0.9	1.5	0.60
80+%	58.0	61.3	0.95	2.4	1.9	1.26	0.0	0.0	...	0.1	0.2	0.50	0.1	0.3	0.33
Total	100.0	100.0		99.9	100.0		100.0	100.0		100.0	100.0		100.0	100.0	

Note. The sample sizes were 1.6 million for KP and 7.4 million for GP. The 22 counties were Alameda, Contra Costa, El Dorado, Fresno, King, Madera, Marin, Napa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara, Santa Cruz, Solano, Sonoma, Stanislaus, Sutter, Tulare, Yolo, and Yuba.
[a]Can be of any race.

TABLE 6—Demographic Comparison of Block-Group Socioeconomic Characteristics of 1985 Kaiser Permanente (KP) Membership and 1980 General Population (GP): 22 Selected Counties in Northern California*

Demographic Variable	Block-Group Composition	Percentage Living in Block Groups with Specified Socioeconomic Composition		KP/GP Ratio
		KP	GP	
Class	<66% working class	39.7	42.4	0.94
	≥66% working class	60.3	57.6	1.05
Poverty	<20% below poverty	91.3	86.8	1.05
	≥20% below poverty	8.7	13.2	0.66
Education	<25% adults less than high school	66.2	61.0	1.09
	≥25% adults less than high school	33.8	39.0	0.87

Note. The sample sizes were 1.6 million for KP and 7.4 million for GP. The 22 counties were Alameda, Contra Costa, El Dorado, Fresno, King, Madera, Marin, Napa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara, Santa Cruz, Solano, Sonoma, Stanislaus, Sutter, Tulare, Yolo, and Yuba.

Several British investigations likewise have found that area-based measures of deprivation and individual-level social class data detect comparable social gradients in health outcomes.[47–53] As noted by Carstairs and Morris,[48] and by Alexander et al.,[50] these area-based measures possess the additional advantage of being able to evaluate the relationship between health status and socioeconomic position among persons not easily classified by traditional occupation-based class measures (e.g., children, students, and adults not in the active labor force). In these studies, area-based measures of deprivation also revealed intraclass mortality gradients,[49–51] providing additional evidence of contextual effects.

Potential Biases Affecting the Census-Based Methodology

The census-based methodology is not without flaws. First, it requires that individuals have a residential address, and that this address be located within a census-defined tract or block group. In this investigation, 17% of the KPMCP membership could not be geocoded to the block-group level. Despite their similarity in age and gender composition, members with and without block-group data may have differed with regard to other sociodemographic characteristics (e.g., urban vs rural residence), thereby introducing bias. Manual searches on addresses not geocoded by computerized procedures, however, can reduce the percentage of persons not assigned a block-group number to under 5%.[14,15,21] Moreover, beginning with the 1990 census, block-group codes were assigned to the entire nation.[54]

Second, census data might not accurately characterize the demographic context of study subjects because of the decennial nature of census data[55] and also because of the undercount, which chiefly affects poor people and people of color.[55–58] Because population growth and migration can alter a neighborhood's composition,[55] this census-based methodology should be used only for residential addresses falling within 5 years of the closest census. The undercount, in turn, most likely would dilute the association between health status and block-group socioeconomic measures, since it produces conservative estimates of the number and hence proportion of poor persons and people of color.[55–58]

A third concern pertains to ecologic fallacy, which can produce inflated estimates of individual-level associations.[22–23] Although this study used grouped data to characterize individuals, the census-based methodology presented here is not affected by the "classic" type of ecologic fallacy, in which both the dependent and independent variables are grouped data and underlying factors associated with the grouping process confound the results.[22–23] Instead, individual-level dependent variables were analyzed in conjunction with census-based independent variables. To minimize heterogeneity in the census unit, however, it may be preferable to use block-group rather than tract data, if available.[21,38–40]

A final caveat applies to the contextual analyses. Only a limited number of socioeconomic and health-related variables were used in these analyses, since the purpose was to evaluate the association between different socioeconomic factors and health characteristics, as opposed to elucidating the pathways through which these factors exert their effects. Yet, as noted by some critics, if individual-level models are not fully specified, seemingly significant contex-

tual effects may be spurious, with the contextual variables only "explaining" variance that could be better accounted for by additional individual-level data.[37,59] Even so, the plausible nature of the observed contextual trends suggests that this technique may provide a useful means by which to avoid "individualistic fallacy,"[22] that is, the assumption that individual-level data are sufficient to explain social phenomena, including a population's level of health and disease.[60]

Significance and Usefulness of the Census-Based Methodology

The importance of validating this census-based approach to measuring socioeconomic position is underscored by the numerous US studies that, in the absence of individual-level social class data, have used census-derived data from people's immediate neighborhoods in conjunction with individual-level health data to describe, analyze, or control for social gradients in various health outcomes.[9–20] These include investigations regarding race/ethnic differences in cancer incidence and survival,[9–16] homicide,[17] and childhood diseases,[18] as well as studies examining intraurban variation in mortality.[19,20] All have observed significant associations between people's health status and the socioeconomic conditions of the neighborhoods in which they live, and all have expressed concerns regarding the use of census-derived data. The results of this study and comparable research[21,47–53] indicate that these prior findings most likely are legitimate and probably underestimate the effect that would have been observed were individual-level social class data available.

In sum, the census-based methodology presented in this study provides a valid and useful approach to overcoming the absence of socioeconomic data in most US medical records. Because census data are readily available to all US public health researchers and geocoding is not expensive, this methodology easily can be used with any existing data set that contains residential addresses. By offering a measure of socioeconomic position that is applicable to all persons, regardless of age, gender, or employment status, this census-based approach also can greatly assist in both reducing and evaluating response bias resulting from incomplete or inaccurate individual-level socioeconomic information. Other possible uses include ascertaining whether a study sample is representative of the general public, whether a given program is reaching its intended population, and whether socioeconomic factors contribute to small-area variation in health status and health service use patterns. Lastly, in those cases where individual-level data are available, this methodology permits use of contextual analysis, thereby offering new avenues to investigate the diverse routes by which social gradients in disease are produced and maintained.

Apart from raising new questions about social context, the most important contribution of the census-based methodology may be in furnishing US researchers with the means to construct age-specific incidence, prevalence, and mortality rates, stratified by consistent area-based measures of social class. This can be accomplished for most disease outcomes, since denominator data typically are census derived and can be classified according to the same block-group measures.[14] By focusing attention on social gradients in disease within, and not only between, different race/ethnic groups within the United States, such data could potentially reinvigorate, if not reorient, efforts to understand as well as eliminate socially determined disparities in health. □

Acknowledgments

This project was funded by National Cancer Institute contract N01-CP-95606.

This paper was presented at the 118th Annual Meeting of the American Public Health Association, October 1990, New York City.

Thanks to Frank Many for help in obtaining the 1980 census data, to Donna Wells for programming assistance, to Bruce Fireman for suggestions regarding the statistical analyses, and to Gary Friedman and Robert A. Hiatt for their helpful comments and criticisms.

References

1. Liberatos P, Link BG, Kelsey JL. The measurement of social class in epidemiology. *Epidemiol Rev.* 1988;10:87–121.
2. US Department of Health and Human Services. *Health Status of Minorities and Low Income Groups.* Washington, DC: US Government Printing Office; 1985. DHHS publication no. (HRSA) HRS-P-DV 85-1.
3. Sydenstricker E. *Health and Environment.* New York, NY: McGraw-Hill; 1933.
4. Susser M, Watson W, Hopper K. *Sociology in Medicine.* 3rd ed. New York, NY: Oxford University Press; 1985.
5. Black D, Morris JN, Smith C, Townsend P. *Inequalities in Health: The Black Report.* Harmondsworth, England: Penguin Books; 1985.
6. Smith GD, Bartley M, Blanc D. The Black report on socioeconomic inequalities in health 10 years on. *Br Med J.* 1990;301:373–377.
7. Marmot MG. Social inequalities in mortality: the social environment. In: Wilkinson RG, ed. *Class and Health: Research and Longitudinal Data.* London, England: Tavistock; 1986:21–33.
8. Morgenstern H. Socioeconomic factors: concepts, measurements, and health effects. In: Ostfeld AM, Eaker ED, eds. *Measuring Psycho-Social Variables in Epidemiologic Studies of Cardiovascular Disease.* Bethesda, Md: National Institutes of Health; 1985:3–35. NIH publication no. 85-2270.
9. Devesa SS, Diamond EL. Association of breast cancer and cervical cancer incidence with income and education among Whites and Blacks. *JNCI.* 1980;65:515–528.
10. Devesa SS, Diamond EL. Socioeconomic and racial differences in lung cancer incidence. *Am J Epidemiol.* 1983;118:818–831.
11. Ernster VL, Selvin S, Sacks ST, et al. Prostatic cancer: mortality and incidence rates by race and social class. *Am J Epidemiol.* 1978;107:311–320.
12. Dayal HH, Power RN, Chiu C. Race and socioeconomic status in survival from breast cancer. *J Chronic Dis.* 1982;35:675–683.
13. White E, Daling JR, Norsted TL, et al. Rising incidence of breast cancer among young women in Washington State. *JNCI.* 1987;79:239–243.
14. Krieger N. Social class and the Black/White crossover in the age-specific incidence of breast cancer: a study linking census-derived data to population-based registry records. *Am J Epidemiol.* 1990;131:804–814.
15. Bassett MT, Krieger N. Social class and Black-White differences in breast cancer survival. *Am J Public Health.* 1986;76:1400–1403.
16. Savage D, Lindenbaum J, Ryzin JV, et al. Race, poverty, and multiple myeloma. *Cancer.* 1984;54:3085–3094.
17. Centerwall BS. Race, socioeconomic status, and domestic homicide, Atlanta, 1971–72. *Am J Public Health.* 1984;74:813–815.
18. Wise PH, Kotelchuck M, Wilson ML, Mills M. Racial and socioeconomic disparities in childhood mortality in Boston. *N Engl J Med.* 1985;316:360–366.
19. Jenkins CD, Tuthill RW, Tannenbaum SI, Kirby CI. Zones of excess mortality in Massachusetts. *N Engl J Med.* 1977;296:1354–1356.
20. Yeracaris CA, Kim JH. Socioeconomic differences in selected causes of death. *Am J Public Health.* 1978;68:342–351.
21. Krieger N. Women and social class: a methodological study comparing individual, household, and census measures as predictors of Black/White differences in reproductive history. *J Epidemiol Community Health.* 1991;45:35–42.
22. Alker HR Jr. A typology of ecological fallacies. In: Doggan M, Rokkan S, eds. *Social Ecology.* Cambridge, Mass: MIT Press; 1969:69–86.
23. Selvin S. Two issues concerning the analysis of grouped data. *Eur J Epidemiol.* 1987;3:284–287.
24. Kaplan CP, Van Valey TL. *Census '80: Continuing the Fact Finding Tradition.* Washington, DC: US Government Printing Office; 1980.
25. Syme SL, Oakes TW, Friedman GD, Feldman R, Siegelaub AB, Collen M. Social

class and racial differences in blood pressure. *Am J Public Health.* 1974;64:619–620.
26. Walker M, Shaper AG, Wannamethee G. Height and social class in middle-aged British men. *J Epidemiol Community Health.* 1988;42:299–303.
27. National Survey of Family Growth. *Socioeconomic Differentials and Trends in the Timing of Births.* Hyattsville, MD: National Center for Health Statistics; 1981. DHHS publication no. (PHS) 81-1982. (Vital and health statistics; series 23; no. 6).
28. Wright EO, Costello C, Hacken D, Sprague J. The American class structure. *Am Sociol Rev.* 1982;47:709–726.
29. US Bureau of the Census. *Poverty Areas in Large Cities. 1980 Census of the Population, Volume 2, Subject Reports.* Washington, DC: US Government Printing Office; 1985. PC80-2-8D.
30. Congressional Budget Office. *Trends in Family Income: 1970–1986.* Washington, DC: US Government Printing Office; 1988.
31. Beeghley L. Illusion and reality in the measurement of poverty. *Soc Problems.* 1984; 31:322–333.
32. Friedman GD, Selby JV, Quesenberry CP Jr, Armstrong MA, Klatsky AL. Precursors of essential hypertension: body weight, alcohol and salt use, and parental history of hypertension. *Prev Med.* 1988;17:387–402.
33. SAS Institute Inc. *SAS Language Guide for Personal Computers.* Release 6.03 edition. Cary, NC: SAS Institute Inc; 1988.
34. Kleinbaum DE, Kupper LL. *Applied Regression Analysis and Other Multivariable Methods.* Boston, Mass: Duxbury Press; 1978.
35. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Volume I—The Analysis of Case-Control Studies.* Lyon, France: International Agency for Research on Cancer; 1980:191–246.
36. Boyd LH Jr, Iversen GR. *Contextual Analysis: Concepts and Statistical Techniques.* Belmont, Calif: Wadsworth Publishing Co; 1979.
37. Blalock HM Jr. Contextual-effects models: theoretic and methodological issues. *Annu Rev Sociol.* 1984;10:353–372.
38. Foley DL. Census tracts and urban research. *J Am Stat Assoc.* 1953;48:733–742.
39. Myers JK. Note on the homogeneity of census tracts: a methodological problem in urban ecological research. *Soc Forces.* 1954;32:364–366.
40. US Bureau of the Census. *Census Use Study: Health Information System—II.* Report no. 12. Washington, DC: US Government Printing Office; 1976.
41. Haan M, Kaplan G, Camacho T. Poverty and health: prospective evidence from the Alameda county study. *Am J Epidemiol.* 1987;125:989–998.
42. Przeworski A. Contextual models of political behavior. *Political Methodology.* 1974;1:27–61.
43. Segal DR, Meyer MW. The social context of political partisanship. In: Doggan M, Rokkan S, eds. *Social Ecology.* Cambridge, Mass: MIT Press; 1969:217–232.
44. Langbein LI. Schools or students: aggregation problems in the study of achievement. *Eval Stud Annu Rev.* 1977;2:270–298.
45. Bell W. Urban neighborhoods and individual behavior. In: Sherif M, Sherif CW, eds. *Problems of Youth: Transition to Adulthood in a Changing World.* Chicago, Ill: Aldine; 1965:235–264.
46. Erbe BM. Race and socioeconomic segregation. *Am Sociol Rev.* 1975;40:801–812.
47. Curtis SE. Use of survey data and small area statistics to assess the link between individual morbidity and neighborhood deprivation. *J Epidemiol Community Health.* 1990;49:62–68.
48. Carstairs V, Morris R. Deprivation: explaining differences in mortality between Scotland and England and Wales. *Br Med J.* 1989;299:886–889.
49. Carstairs V, Morris R. Deprivation and mortality: an alternative to social class? *Community Med.* 1989;11:210–219.
50. Alexander FE, O'Brien F, Hepburn W, Miller M. Association between mortality among women and socioeconomic factors in general practices in Edinburgh: an application of small area statistics. *Br Med J.* 1987;295:754–756.
51. Morgan M, Chinn S. ACORN group, social class, and child health. *J Epidemiol Community Health.* 1983;37:196–203.
52. Morgan M. Measuring social inequality: occupational classifications and their alternatives. *Community Med.* 1983;5:116–124.
53. Arber S. Social class, nonemployment, and chronic illness: continuing the inequalities in health debate. *Br Med J.* 1987;294:1069–1073.
54. US Bureau of the Census. *Census '90 Basics.* Washington, DC: US Government Printing Office; 1990.
55. White MJ. *American Neighborhoods and Residential Differentiation.* New York, NY: Russell Sage Foundation; 1987.
56. Heer DM, ed. *Social Statistics and the City.* Cambridge, Mass: Joint Center for Urban Studies at the Massachusetts Institute of Technology and Harvard University; 1968.
57. Parsons CW, ed. *America's Uncounted People: Report of the Advisory Committee on Problems of Census Enumeration, Division of Behavioral Sciences, National Research Council.* Washington, DC: US Government Printing Office; 1972.
58. Citro F, Cohen ML, eds. *The Bicentennial Census: New Directions for Methodology in 1990.* Washington, DC: National Academy of Sciences; 1985.
59. Hauser RM. Context and consex: a cautionary tale. *Am J Sociol.* 1970;75:645–664.
60. Rose G. Sick individuals and sick populations. *Int J Epidemiol.* 1985;14:32–38.