Zip Code Caveat: Bias Due to Spatiotemporal Mismatches Between Zip Codes and US Census–Defined Geographic Areas—The Public Health Disparities Geocoding Project

Nancy Krieger, PhD, Pamela Waterman, MPH, Jarvis T. Chen, ScD, Mah-Jabeen Soobader, PhD, S. V. Subramanian, PhD, and Rosa Carson, BA

Use of zip codes in US public health research is on the rise. As of February 2002, 230 articles were indexed by zip code in PubMed,[1] all published since 1989. Fifty-two of these articles (23%) involved the use of census-derived zip code socioeconomic data (e.g., median household income) to investigate the effects of socioeconomic position on specified health outcomes (article citations are available on request from the authors).

To date, discussions regarding the use of zip code socioeconomic data for US public health research have focused chiefly on whether zip codes' larger population size (average: 30 000) and potentially greater socioeconomic heterogeneity would attenuate estimates of socioeconomic gradients in health detected using zip codes in comparison with estimates obtained via census tract (average population: 4000) or block group (average population: 1000) socioeconomic data.[2–7] Unacknowledged in the public health literature, however, is the fact that zip codes differ from census tracts and block groups in other important ways, including spatiotemporal definition and stability.

Unlike census tracts, defined by the US Bureau of the Census as "small, relatively permanent statistical subdivision[s] of a county . . . designed to be relatively homogeneous with respect to population characteristics, economic status, and living conditions,"[8(ppG-10–G-11)] zip codes are "administrative units established by the United States Postal Service . . . for the most efficient delivery of mail, and therefore generally do not respect political or census statistical area boundaries."[9(pA-13)] Spanning in size from a single building or company with a high volume of mail to large areas that cut across states, "carrier routes for one zip code may intertwine with those of one or more zip codes" such that "this area is more conceptual than geographic."[10(p22)]

To "overcome the difficulties in precisely defining the land area covered by each zip code,"[11] the US Census Bureau created a new statistical entity built from census blocks: the 5-digit zip code tabulation area (ZCTA), first used in the 2000 census.[12] Of note, ZCTAs and zip codes sharing the same 5-digit code may not necessarily cover the same area (Table 1),[13] so that zip codes obtained via self-report or from addresses in medical records cannot be assumed to correspond to census-defined ZCTAs.

Even before introduction of the ZCTAs, there were 2 types of spatiotemporal discontinuity that could conceivably affect health studies linking zip codes to census-derived data: (1) changes in zip code delivery routes—and hence in population covered by the affected zip code—and (2) discontinuation and addition of zip codes in nondecennial years.[14–16] Between 1997 and 2001 alone, the US Post Office added approximately 390 new zip codes nationwide and discontinued 120 (oral communication, Meg Ausman, US Post Office Data Center, February 5, 2002). One implication of these changes is that persons could be correctly geocoded to a zip code that did not exist in the preceding decennial census.

Findings from the Public Health Disparities Geocoding Project[17] illustrate the potential problems for health research of spatiotemporal zip code–census mismatches, even those dating from before the creation of ZCTAs. This project was designed to assess which area-based socioeconomic measures at which levels of geography (census tract, block group, and zip code) are most appropriate for monitoring socioeconomic inequalities in health. Health data from 2 states (Massachusetts and Rhode Island) and the 1990 census were used. Records were geocoded in 1999 by a firm whose accuracy we ascertained to be high (96%),[18] and the firm, following standard practice, returned the most recent geocodes available.

TABLE 1—Technical Definitions of and Distinctions Between Zip Codes and Zip Code Tabulation Areas (ZCTAs)

Definition of ZCTAs[11]

"ZIP Code Tabulation Areas (ZCTAs™) are a new statistical entity developed by the US Census Bureau for tabulating summary statistics from Census 2000. This new entity was developed to overcome the difficulties in precisely defining the land area covered by each ZIP Code®. Defining the extent of an area is necessary to accurately tabulate census data for that area. ZCTAs are generalized area representations of US Postal Service (USPS) ZIP Code service areas. Simply put, each one is built by aggregating the Census 2000 blocks, whose addresses use a given ZIP Code, into a ZCTA which gets that ZIP Code assigned as its ZCTA code. They represent the majority USPS five-digit ZIP Code found in a given area. For those areas where it is difficult to determine the prevailing five-digit ZIP Code, the higher-level three-digit ZIP Code is used for the ZCTA code. Since we take the ZIP Code used by the majority of addresses in an area for the ZCTA code, some addresses will end up with a ZCTA code different from their ZIP Code. Also, some ZIP Codes represent very few addresses (sometimes only one) and therefore will not appear in the ZCTA universe."

Distinction between ZCTAs and Zip Codes[12]

"Even though the codes may appear the same, the addresses and areas covered by these areas may not be the same. We strongly advise data users who wish to compare 1990 and 2000 data to determine and evaluate any coverage differences that exist before making any comparisons. There are several reasons for this caution: The USPS has extensively modified ZIP Codes over the last ten years. Even though a 1990 ZIP Code matches a Census 2000 ZCTA code, there is no guarantee that these cover the same geographic area. Also, some ZIP Codes in the 1990 data products were discontinued by the USPS, and new ZIP Codes were created; ZCTAs and the 1990 data products were discontinued by the USPS, and new ZIP Codes were created; ZCTAs and the 1990 census ZIP Code areas were delineated using different methodologies and therefore may not have comparable coverage area or size; and the Census 2000 ZCTAs will include some dedicated PO box ZIP Codes. All dedicated PO box ZIP Codes were excluded as ZIP Code areas in 1990. The resulting 1990 areas include data for both PO box ZIP Codes and the ZIP Codes that provide street or rural route delivery to the surrounding area."

TABLE 2—Incident Colon Cancer Counts by Geographic Level: Massachusetts, 1987–1993

	Geocoded to			
			Zip Code	
No. of Cases of Colon Cancer	Block Group No. (%)	Census Tract, No. (%)	Total, No. (%)	Zip Code Changed or Established After the 1990 Census, No. (%)
17 266	15 792 (91.5)	17 265 (100.0)	17 266 (100.0)	1784 (10.3)

Cancer incidence rates were one of the health outcomes addressed. We found that in Massachusetts (474 zip codes listed in the 1990 census), 17 376 (10.4%) of the 166 730 cancer cases occurring during 1987 to 1993 were geocoded to 193 zip codes not included in the 1990 census; 15 774 (90.8%) of these 17 376 cases were in one of 30 zip codes changed or established after the 1990 census.[19–21] By contrast, in Rhode Island (70 zip codes listed in the 1990 census), only 0.7% (148) of the 19 766 geocoded cancer incidence records were matched to zip codes not included in the 1990 census.

In the case of colon cancer incidence in Massachusetts, moreover, the impact of excluding persons linked to zip codes not included in the 1990 census was substantial. Zip code–level analyses yielded socioeconomic gradients contrary to those observed via data at the tract and block group levels and contrary to those reported in the literature (Tables 2–4).[22]

Given the growing interest in linking geographic and health data,[23,24] we urge researchers, when using geocoded records, to pay careful attention to the potential for spatiotemporal mismatches between census-derived and zip code data as well as to changes in zip code boundaries and years in which boundaries were established. Public health projects and programs that use zip code data should likewise be alert to potential new issues stemming from the replacement of zip codes with ZCTAs in the 2000 census. ■

About the Authors

The authors are with the Department of Health and Social Behavior, Harvard School of Public Health, Boston, Mass.

Requests for reprints should be sent to Nancy Krieger, PhD, Department of Health and Social Behavior, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115 (e-mail: nkrieger@hsph.harvard.edu).

This brief was accepted March 15, 2002.

Contributors

N. Krieger conceived and designed the study, directed data analysis, interpreted the data, and wrote the article. P. Waterman, J.T. Chen, M. Soobader, and S.V. Subramanian contributed to the conception and design of the study and analyzed and assisted with interpretation of the data. R. Carson assisted with presentation of the study results.

Acknowledgments

This work was funded by the National Institute of Child Health and Human Development, National Institutes of Health (grant 1 R01 HD36865-01).

We thank Dr Daniel Friedman (assistant commissioner, Bureau of Health Statistics, Research and Evaluation, Massachusetts Department of Health) and Dr Jay Buechner (chief, Office of Health Statistics, Rhode Island Department of Health) for facilitating the Public Health Disparities Geocoding Project by providing data from their respective health departments and for their helpful comments on the article. We likewise thank the following individuals for helping us to access the mortality and cancer registry data: Alice Mroszczyk, Dr Susan Gershman, Mary Mroszczyk, and Ann R. MacMillan of the Massachusetts Department of Public Health and Dr John Fulton of the Rhode Island Department of Health.

References

1. National Library of Medicine. PubMed. Available at: http://www.ncbi.nlm.nih.gov/PubMed/. Accessed February 5, 2002.

2. Krieger N, Williams D, Moss N. Measuring social class in US public health research: concepts, methodologies and guidelines. *Annu Rev Public Health.* 1997;18:341–378.

3. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *Am J Public Health.* 1992;82:703–710.

4. Geronimus AT, Bound J. Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. *Am J Epidemiol.* 1998;48:475–486.

5. Diez Roux AV, Kiefe CI, Jacobs DR Jr, et al. Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies. *Ann Epidemiol.* 2001;11:395–405.

6. Fiscella K, Franks P. Impact of patient socioeconomic status on physician profiles: a comparison of census-derived and individual measures. *Med Care.* 2001;39:8–14.

TABLE 3—Colon Cancer Incidence Rates, Stratified by Area-Based Socioeconomic Measures, Among Persons in Areas With the Least and Most Resources, Along With Age-Adjusted Comparisons (Incidence Rate Ratio and Relative Index of Inequality): Massachusetts, 1987–1993

Selected Area-Based Socioeconomic Measure[a]	Rate: Least Resources[b]			Rate: Most Resources[b]			Incidence Rate Ratio (95% Confidence Interval): Least/Most			Relative Index of Inequality (95% Confidence Interval)		
	BG	CT	ZC	BG	CT	ZC	BG	CT	ZC	BG	CT	ZC
Working class (categorical)	41.3	42.5	41.1	45.8	48.3	27.9	0.90 (0.76, 1.06)	0.88 (0.73, 1.06)	1.47 (1.14, 1.90)	0.89 (0.84, 0.95)	0.85 (0.80, 0.90)	1.28 (1.20, 1.36)
Median household income (quintile)	41.0	42.5	42.3	46.3	48.9	37.2	0.89 (0.75, 1.04)	0.87 (0.74, 1.03)	1.14 (0.97, 1.34)	0.87 (0.82, 0.93)	0.88 (0.83, 0.93)	1.19 (1.12, 1.27)
Poverty (categorical)	41.7	45.6	44.8	43.9	47.4	41.6	0.95 (0.80, 1.13)	0.96 (0.81, 1.15)	1.08 (0.88, 1.32)	0.94 (0.88, 1.00)	0.95 (0.89, 1.01)	1.06 (0.99, 1.13)
Low education (categorical)	39.5	40.8	43.8	45.2	48.0	39.3	0.87 (0.73, 1.05)	0.85 (0.70, 1.03)	1.11 (0.90, 1.38)	0.84 (0.79, 0.90)	0.90 (0.85, 0.96)	1.15 (1.08, 1.22)
Index of local economic resources (quintile)	40.3	42.6	43.1	45.4	48.7	33.6	0.89 (0.76, 1.04)	0.87 (0.74, 1.03)	1.28 (1.09, 1.50)	0.86 (0.81, 0.91)	0.88 (0.83, 0.94)	1.27 (1.19, 1.35)

Note. The relative index of inequality is a measure of effect that takes into account both the population distribution of the exposure and the magnitude of the rate ratio detected in each socioeconomic stratum, thereby permitting meaningful comparison of gradients across different socioeconomic measures.[25–27] BG = block group; CT = census tract; ZC = zip code.

[a]The area-based socioeconomic measures and their cutpoints for these analyses are defined in Table 4.[17]

[b]Average annual rate (per 100 000) age standardized to the year 2000 standard million.[28]

TABLE 4—Area-Based Socioeconomic Measures and Cutpoints Used in Data Analysis

Selected Area-Based Socioeconomic Measure	Operational Definition and Cut Points Used
Working class[2] (categorical)	Percentage of persons employed in predominantly working class occupations (i.e., as nonsupervisory employees), operationalized as percentage of persons employed in the following 8 of 13 census-based occupational groups: administrative support; sales; private household service; other service (except protective); precision production, craft, repair; machine operators, assemblers, inspectors; transportation and material moving; handlers, equipment cleaners, laborers; cutpoints: C1 = 0%–49.9%, C2 = 50%–69.9%, C3 = 66%–74.9%, C4 = 75%–100%
Median household income (quintile)	Median household income in year before the decennial census (US in 1989: $30 056); cutpoints:
	Massachusetts BG: Q1 = $4999–$26 110, Q2 = $26 111–$33 749, Q3 = $33 750–$40 798, Q4 = $40 799–$49 903, Q5 = $49 904–$150 001
	Massachusetts CT: Q1 = $4999–$26 471, Q2 = $26 472–$33 162, Q3 = $33 163–$39 286, Q4 = $39 287–$47 124, Q5 = $47 125–$102 797
	Massachusetts ZC: Q1 = $9726–$30 624, Q2 = $30 625–$36 246, Q3 = $36 247–$41 396, Q4 = $41 397–$48 841, Q5 = $48 842–$94 898
Poverty (categorical)	Percentage of persons below federally defined poverty line, a threshold that varies by size and age composition of the household and, on average, equaled $12 647 for a family of 4 in 1989[9]; cutpoints: C1 = 0%–4.9%, C2 = 5.0%–9.9%, C3 = 10.0%–19.9%, C4 = 20%–100%; areas with a poverty rate of ≥ 20% are federally defined poverty areas[2]
Low education (categorical)	Percentage of persons 25 years and older with less than a 12th grade education; cutpoints: C1 = 0%–14.9%, C2 = 15.0%–24.9%, C3 = 25.0%–39.9%, C4 = 40%–100%
Index of local economic resources[29] (quintile)	A "summary index" used by the Centers for Disease Control and Prevention and based on "white collar employment, unemployment, and family income"; cutpoints:
	Massachusetts BG: Q1 = 0–6, Q2 = 7–11, Q3 = 12–15, Q4 = 16–20, Q5 = 21–27
	Massachusetts CT: Q1 = 0–5, Q2 = 6–10, Q3 = 11–15, Q4 = 16–19, Q5 = 20–26
	Massachusetts ZC: Q1 = 0–8, Q2 = 9–12, Q3 = 13–15, Q4 = 16–19, Q5 = 20–26

Note. C = category; BG = block group; Q = quintile; CT = census tract; ZC = zip code.

7. Diez Roux AV. Investigating neighborhood and area effects on health. *Am J Public Health.* 2001;91:1783–1789.

8. US Bureau of the Census. Geographical areas reference manual. Available at: http://www.census.gov/geo/www/garm.html. Accessed February 12, 2002.

9. *Census of Population and Housing, 1990: Summary Tape File 3 Technical Documentation.* Washington, DC: US Bureau of the Census; 1991.

10. US Bureau of the Census. Geographics changes for census 2000 + glossary. Available at: http://www.census.gov/geo/www/tiger/glossary.html. Accessed July 3, 2001.

11. US Bureau of the Census. Census 2000 zip code tabulation areas (ZCTAs). Available at: http://www.census.gov/geo/ZCTA/zcta.html. Accessed February 12, 2002.

12. US Bureau of the Census. Zip code tabulation area (ZCTA) frequently asked questions. Available at: http://www.census.gov/geo/ZCTA/zctafaq.html. Accessed February 12, 2002.

13. US Bureau of the Census. Census 2000 zip code tabulation areas technical documentation. Available at: http://www.census.gov/geo/ZCTA/zcta_tech_doc.pdf. Accessed February 12, 2002.

14. *Address Information Products Technical Guide.* Memphis, Tenn: US Postal Service National Customer Support Center; 2001.

15. Office of Social and Economic Data Analysis. The zip code resource page: tools and resources related to US zip codes. Available at: http://www.oseda.missouri.edu/uic/zip.resources.html. Accessed: February 24, 2002.

16. Range L, Clay T, Oliva G. *California Child and Youth Injury Hot Spot Project 1995–1997, Volume 3: Technical Guide.* Sacramento, Calif: California Dept of Health Services, Maternal and Child Health Branch; 2000:59–67.

17. Krieger N, Chen JT, Waterman PD, Soobader M-J, Subramanian SV, Carson R. Geocoding and monitoring US socioeconomic inequalities in mortality and cancer incidence: does choice of area-based measure and geographic level matter?—the Public Health Disparities Geocoding Project. *Am J Epidemiol.* In press.

18. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating accuracy of geocoding for public health research. *Am J Public Health.* 2001;91:1114–1116.

19. *National Five-Digit Zip Code and Post Office Directory.* Memphis, Tenn: US Postal Service; 1993:8-3. Publication 65.

20. *National Five-Digit Zip Code and Post Office Directory.* Memphis, Tenn: US Postal Service; 2000:3-1328–3-1371. Publication 65.

21. *Zip Code Changes for the Boston Area Effective July 1, 1998.* Dorchester, Mass: US Post Office; 1998.

22. Kogevinas M, Pearce N, Susser M, Bofetta P, eds. *Social Inequalities in Cancer.* Lyon, France: International Agency for Research on Cancer; 1997. IARC scientific publication 136.

23. Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. Geographic information systems and public health: mapping the future. *Public Health Rep.* 1999; 114:359–373.

24. Moore DA, Carpenter TE. Spatial analytic methods and geographic information systems: use in health research and epidemiology. *Epidemiol Rev.* 1999;21: 143–161.

25. Pamuk ER. Social class inequality in mortality from 1921 to 1972 in England and Wales. *Popul Stud.* 1985;39:17–31.

26. Wagstaff A, Paci P, van Doorslaer E. On the measurement of inequalities in health. *Soc Sci Med.* 1991; 33:545–557.

27. Davey Smith G, Hart C, Hole D, et al. Education and occupational social class: which is the more important indicator of mortality risk? *J Epidemiol Community Health.* 1998;52:153–160.

28. Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard. *Natl Vital Stat Rep.* October 7, 1998;47(3).

29. Casper ML, Barnett E, Halverson JA, et al. *Women and Heart Disease: An Atlas of Racial and Ethnic Disparities in Mortality.* Morgantown, WVa: Office for Social Environment and Health Research, West Virginia University; 1999.