# The Public Health Disparities Geocoding Project Monograph (June 2004)

**The Public Health Disparities Geocoding Project Monograph**

These pages present an introduction to geocoding and using area-based socioeconomic measures with public health surveillance data, based on the work of the Public Health Disparities Geocoding Project at the Harvard T. H Chan School of Public Health, Department of Social and Behavioral Sciences.

• The **Executive Summary** describes the motivation behind the Public Health Disparities Geocoding Project, and summarizes the methodology, key findings, and recommendations.

• The **Introduction** provides a more in-depth look at the history of geocoding and area-based measures, the objectives of our project, and our main findings. We include a glimpse of what routine public health surveillance of socioeconomic disparities in health could look like if conducted over a variety of health outcomes over the lifecourse, from birth to death, using a single area-based socioeconomic measure at the census tract level.

• The **Publications** page is a comprehensive list of the publications of the Public Health Disparities Geocoding Project, and includes pdf copies of all of our published work.

• We also provide a primer on the basics of **Geocoding**, including descriptions of the many options and services available, and the nitty-gritty details of address cleaning, address formatting, and evaluation of geocoding accuracy.

• In **Generating ABSMs** we describe the concepts, methods, and measures behind creating area-based socioeconomic measures, including a summary table of the 19 theoretically justified area-based socioeconomic measures we created based on 1990 U.S. Census data (see ABSM Creation Table).

• Under **Analytic Methods**, we provide details on how to merge geocoded surveillance data with Census derived population denominators and area-based socioeconomic measures. We also present basic epidemiologic methods for generating descriptive statistics, including directly age-standardized incidence rates, incidence rate ratios and rate differences, the relative index of inequality, and population attributable fraction. Examples are provided for each of these techniques, and each section is further detailed in our comprehensive Case Example.

• We've also included some information about **Multi-level Modeling** and **Visual Display** of data for surveillance reporting.

• The **Case Example** is an opportunity for programmers and data managers to try out the techniques we describe on a test dataset, drawn from all-cause mortality cases in Suffolk County, MA, from 1989 to 1991. We provide test datasets, a step-by-step description of the programming tasks, sample SAS code, and examples of the resulting output.

• Finally, to facilitate further research on socioeconomic gradients in health with respect to our recommended area-based socioeconomic measure (CT poverty), we have made available **Census Tract Level Poverty Data** for ALL census tracts in the United States, for 1980, 1990, and 2000.

*Please cite as:*
*Krieger N, Waterman PD, Chen JT, Rehkopf DH, Subramanian SV. The Public Health Disparities Geocoding Project Monograph. Available as of June 30, 2004 at: http://www.hsph.harvard.edu/thegeocodingproject*

# Executive Summary

**The problem**
A lack of socioeconomic data in most US public health surveillance systems.

**Why is this a problem?**
Absent these data, we cannot: (a) monitor socioeconomic inequalities in US health; (b) ascertain their contribution to racial/ethnic and gender inequalities in health; and (c) galvanize public concern, debate, and action concerning how we, as a nation, can achieve the vital goal of eliminating social disparities in health (Healthy People 2010 overarching objective #2).

**Possible solution**
Geocoding public health surveillance data and using census-derived area-based socioeconomic measures (ABSMs) to characterize both the cases and population in the catchment area, thereby enabling computation of rates stratified by the area-based measure of socioeconomic position.

**Knowledge gaps**
Unknown which ABSMs, at which level of geography, would be most apt for monitoring US socioeconomic inequalities in health, overall and within diverse racial/ethnic-gender groups.

**Methodologic study: The Public Health Disparities Geocoding Project**
We accordingly launched the Public Health Disparities Geocoding Project to ascertain which ABSMs, at which geographic level (census block group [BG], census tract [CT], or ZIP Code [ZC]), would be suitable for monitoring US socioeconomic inequalities in the health. Drawing on 1990 census data and public health surveillance systems of 2 New England states, Massachusetts and Rhode Island, we analyzed data for: (a) 7 types of outcomes: mortality (all cause and cause-specific), cancer incidence (all-sites and site-specific), low birth weight, childhood lead poisoning, sexually transmitted infections, tuberculosis, and non-fatal weapons-related injuries, and (b) 18 different ABSMs. We conducted these analyses for both the total population and diverse racial/ethnic-gender groups, at all 3 geographic levels.

**Key findings**
Our key methodologic finding was that the ABSM most apt for monitoring socioeconomic inequalities in health was the census tract (CT) poverty level, since it: (a) consistently detected expected socioeconomic gradients in health across a wide range of health outcomes, among both the total population and diverse racial/ethnic-gender groups, (b) yielded maximal geocoding and linkage to area-based socioeconomic data (compared to BG and ZC data), and (c) was readily interpretable to and could feasibly be used by state health department staff.Using this measure, we were able to provide evidence of powerful socioeconomic gradients for virtually all the outcomes studied, using a common metric, and further demonstrated that: (a) adjusting solely for this measure substantially reduced excess risk observed in the black and Hispanic compared to the white population, and (b) for half the outcomes, over 50% of cases overall would have been averted if everyone's risk equaled that of persons in the least impoverished CT, the only group that consistently achieved Healthy People 2000 goals a decade ahead of time.

**Recommendation**
US public health surveillance data should be geocoded and routinely analyzed using the CT-level measure "percent of persons below poverty," thereby enhancing efforts to track—and improve accountability for addressing—social disparities in health.

**State Health Departments that have issued reports using the methodology of the Public Health Disparities Geocoding Project**

- "The Health of Washington State Supplement: a statewide assessment addressing health disparities by race, ethnic group, poverty and education." September 2004.
- The 2008 Virginia Health Equity Report.
- For a related Canadian analyses, based on 1991 Census of Canada data and deaths from June 4, 1991 to December 31, 2001: Pampalon R, Hamel D, Gamache P. A comparison of individual and area-based socio-economic data for monitoring social inequalities in health.

# Introduction

Making visible the invisible: A new tool for US health departments to monitor – and boost efforts to address – socioeconomic inequalities in health

## The problem: scant socioeconomic data in US public health surveillance systems

Social inequality kills. It unduly deprives individuals and communities experiencing social deprivation of their health, increases their burden of disability and disease, and cuts short their lives [1 – 4]. Recognizing the powerful toll of social inequality on health and well-being, the objectives of Healthy People 2010 seek "to achieve two overarching goals [5]":

- Increase quality and years of healthy life
- Eliminate health disparities

At issue are "health disparities among segments of the population, including differences that occur by gender, race or ethnicity, education or income, disability, geographic location, or sexual orientation [5]."

Yet, despite widespread recognition of the toll of economic deprivation on health, in the US we face a critical problem hampering public health departments' ability to mobilize public concern and resources to eliminate socioeconomic inequalities in health. Why?

The problem is a lack of routine community-based data on the magnitude and trends of socioeconomic inequalities in health, due to the lack of socioeconomic data in most US public health surveillance systems, other than birth and death [6 – 7]. Although specialized surveys, such as the National Health Interview Survey and the Behavioral Risk Factor Surveillance System do collect socioeconomic data, the vast majority of "disease- and condition-specific surveillance systems and administrative data systems do not collect such data [7, pp.18 – 19]." The net effect is to obscure socioeconomic gradients in health and the contribution of economic deprivation to racial/ethnic and gender inequalities in health, at the national, state, and local level [7 – 11].

Rendered invisible, these preventable disparities in health remain hidden to the view of the public and policy-makers alike. The old adage applies: "if you don't ask, you don't know, and if you don't know, you can't act.[8]" Inertia and fatalism flourish, with anecdotal knowledge about "the poor are always sicker and always with us" unchallenged by evidence that the patterning of socioeconomic inequalities in health varies by time and place and hence is not an immutable or unalterable "fact" beyond the reach of concerted effort to change[2, 3, 12 – 14].

The absence of state and local public health surveillance data on socioeconomic inequalities in health has national ramifications. Reflecting the absence of these data, the federal report Health United States 2002 [15], lacked socioeconomic data in 85% of its 71 tables on "Health Status and Determinants;" virtually all of these tables, however, were stratified by "sex, race, and Hispanic origin." Similarly, fully 70% of the 467 U.S. Healthy People 2010 objectives have no socioeconomic targets, given a lack of baseline data [5]. As a nation, we cannot assess whether socioeconomic inequalities are diminishing or growing over time, or if patterns vary by region or state, or by racial/ethnic-gender group, within and across diverse outcomes.

Why does this matter? Because health statistics accurately depicting the population burden of disease, disability and death, as cogently stated in the new federal report Shaping a Health Statistics Vision for the 21st century [7, p.2], "fulfill essential functions for public health, the health services system, and our society". They help us

understand "where we stand in terms of health as individuals, as subgroups, and as a society," including with regard to "the existence of health disparities [7, pp.2 – 3]."

Additionally,

> "Health statistics provide us with the information upon which we can base important public decisions at the local, state, and national levels. Once we have made those public decisions, health statistics make us accountable for the decisions that we have made. Health statistics thus enable us to evaluate the impact of health policies and health programs on the public's health. In short, health statistics give us the information we need to improve the population's health and to reduce health disparities [7, pp.2 – 3]".

Indeed, the critical importance of documenting the social patterning of disease and death has been recognized since the rise of the public health movement in the mid-19th century14 and is of national and global significance [7, 16]. As Edgar Sydenstricker noted, when establishing the first US population-based morbidity studies in the 1920s, these data are crucial to "give glimpses of what the sanitarian has long wanted to see – a picture of the public-health situation as a whole, drawn in proper perspective and painted in true colors [17, p.280]." It was similarly Sydenstricker's profound recognition of the importance of economic deprivation in shaping population health that led to his conducting, in 1935-1936, the first national, federally-sponsored 10-city study on the health impact of the Depression, forerunner to what ultimately became the National Health Interview Survey [18, 19].

Perhaps the most potent reason why it matters to document and monitor socioeconomic disparities in health is that this evidence is vital to boost efforts to reduce these disparities [2, 3, 8 – 14, 16 – 18]. In 1905, Hermann M. Biggs (1859-1923), internationally renowned for his work in the New York City Health Department and later as Commissioner of Health for New York State, roundly declared [20, p.120]:

> "Public health is purchasable. Within natural limitations a community can determine its own death rate."

Biggs' central point was that societal resources, wisely invested, were key to improving population health – and that these resources could only be secured if fundamental data on population health and its determinants were widely understood and appreciated, by the general public and policy-makers alike. Absent data on the public's health, as Biggs and other public health leaders of his generation had learned [14, 20, 21], appeals for resources and regulations to improve the public's health – and for collaboration across different government agencies to develop and implement the necessary policies – would have no standing or clout.

In 1911, the motto "Public Health is Purchasable" became the official slogan of the Monthly Bulletin of the NYC Health Department, with the rationale solidly explained in an editorial by Biggs, reflecting the era's language of social reform [20, p.320]:

> "Disease is a largely removable evil. It continues to afflict humanity, not only because of incomplete knowledge of its causes and lack of individual and public hygiene, but also because it is extensively fostered by harsh economic and industrial conditions and by wretched housing in congested communities. These conditions and consequently the disease which spring from them can be removed by better social organization. No duty of society, acting through its government agencies, is paramount to this obligation to attack the removable cause of disease. The duty of leading this attack and bringing home to public opinion the fact that the community can buy its own health protection is laid upon all health officers, organization and individuals interested in public health movements. For the provision of more and better facilities for the protection of the public health must come in the last analysis through the

> the education of public opinion so that the community shall vividly realize both its needs and its powers. The modern spirit of social religion, dealing with the concrete facts of life, demands the reduction of the death rate as the first result of its activity. The reduction of the death rate is the principal statistical expression and index of human and social progress. It means the saving and lengthening of the lives of thousands of citizens, the extension of the vigorous working period into old age, and the prevention of inefficiency, misery, and suffering. These advances can be made by organized social reform. Public health is purchasable."

Indeed, as suggested by the population health model articulated in Shaping a Health Statistics Vision for the 21st century [7, p.9](Figure 1 below), it is obvious that the field of public health cannot, by itself, improve health and prevent disease; a societal effort is required. As part of this effort, however, it is our singular task—and fundamental responsibility–to provide the data on population distributions of health, disease, disability and death, and social disparities in these outcomes. Or, as stated in Healthy People 2010 [5]:

> "Healthy People 2010 recognizes that communities, States, and national organizations will need to take a multidisciplinary approach to achieving health equity—an approach that involves improving health, education, housing, labor, justice, transportation, agriculture, and the environment, as well as data collection itself. In fact, current data collection methods make it impossible to assess accurately the health status for some populations, particularly relatively small ones."

Improvements in US health over the course of the 20th century, and especially the decline in childhood infectious disease, demonstrate the salience of Biggs' words. So too does a burgeoning European and Canadian literature on the vital necessity of documenting social inequalities in health as an essential component of what policy makers need to take up these disparities a matter of key importance requiring intersectoral work.

Examples of population health reports emphasizing social inequalities in health that galvanized policy initiatives to address these disparities: Canada and the United Kingdom

Population health reports emphasizing social inequalities in health:

- Health Canada. Achieving Health for All: A Framework for Health Promotion (1986).
- Canada Health Canada. Population health/Santé de la Population.
- Health Canada. Toward a Healthy Future – Second Report on the Health of Canadians (1999).

- DHHS (Department of Health and Society Security). Inequalities in Health: Report of a Working Group. London: DHHS, 1980. ("The Black Report"); see also: Townsend P, Davidson N (eds). Inequalities in Health: The Black Report (3rd ed); Whitehead M. The Health Divide. London: Penguin Books, 1988.
- Drever F, Whitehead M (eds). Health Inequalities: Decennial Supplement. London: The Stationary Office, 1997.
- Acheson D, Barker D, Chambers J, Graham H, Marmot M, Whitehead M. The Report of the Independent Inquiry into Health Inequalities. London: The Stationary Office, 1998. ("The Acheson Report")

Subsequent policy initiatives galvanized by these reports:

- Health Canada. Population Health Mobilization: A Regional Strategy – June 1999.

- Health Canada. Strategies for Population Health: Investing in the Health of Canadians. Prepared by the Federal, Provincial and Territorial Advisory Committee on Population Health for the Meeting of Ministers of Health, Halifax, Nova Scotia, Sept 14-15, 1994.

- UK Department of Health. Saving Lives: Our Healthier Nation. London: The Stationary Office, 1999.
- Department of Health. Reducing Health Inequalities: An Action Report. London: Department of Health, 1999.
- UK Department of Health. Our Healthier Nation.

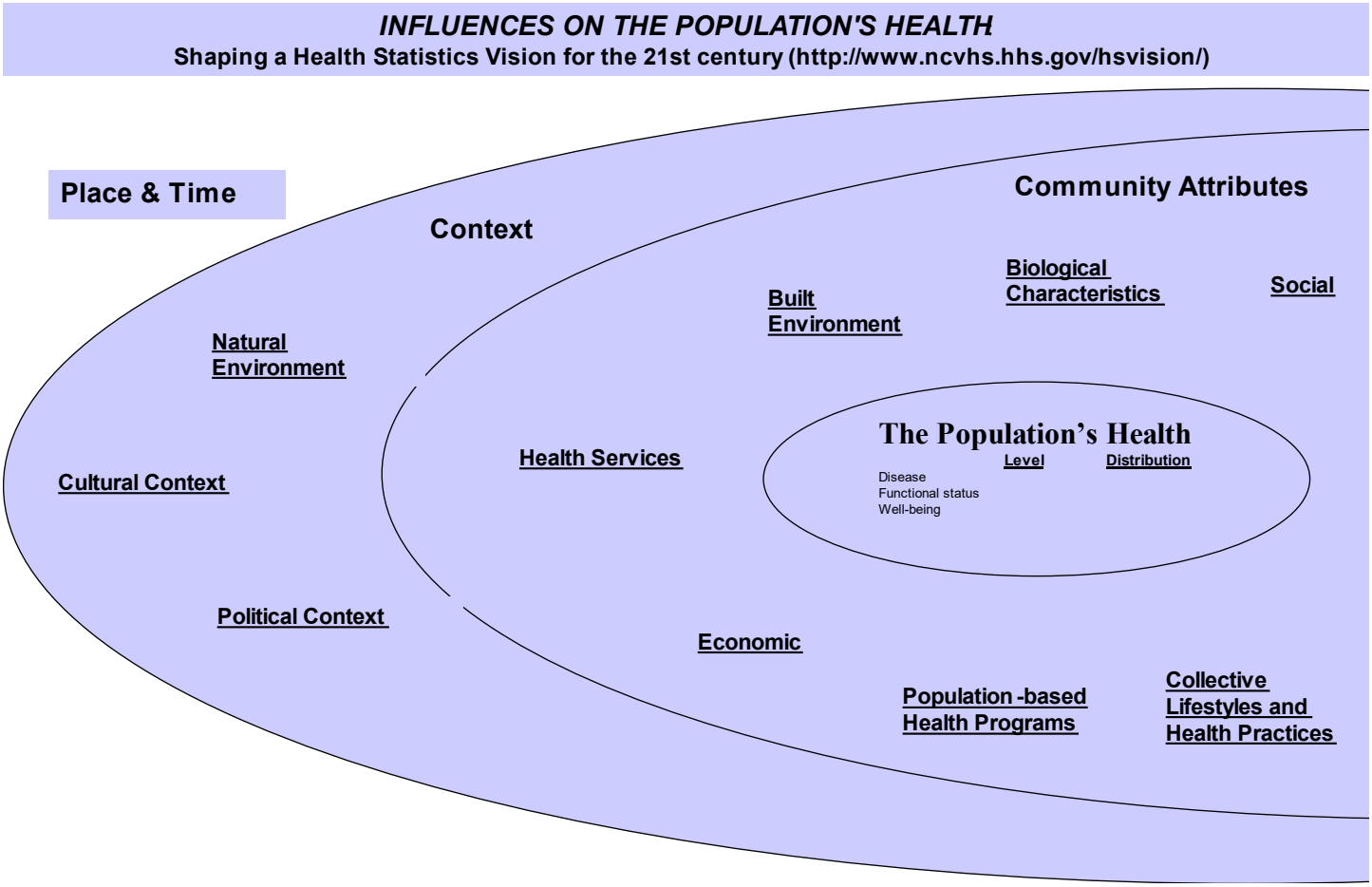



Figure 1: "Influences on the population's health" (from Shaping a Vision of Health Statistics for the 21st Century) (7)

# A solution: geocoding and using area-based socioeconomic measures – key findings of The Public Health Disparities Geocoding Project

Fortunately, one potential and relatively inexpensive solution to the problem of absent or limited socioeconomic data in US public health surveillance systems is provided by the methodology of geocoding residential addresses and using area-based socioeconomic measures (ABSMs) [22 – 26]. In this approach, which draws on multilevel frameworks and area-based measures, both cases (numerators) and the catchment population (denominators) are classified by the socioeconomic characteristics of their residential area, thereby permitting calculation of rates stratified by the ABSMs.

Yet, although this approach has been employed in US health research for over 75 years (see below) [27 – 30], to date there exists no consensus or standard as to which ABSMs, at which level of geography, are best suited for monitoring US socioeconomic inequalities in health, whether within the total population or within diverse racial/ethnic-gender groups [22, 31]. Instead, published research has exhibited a remarkable eclecticism regarding choice of geographic level and types of ABSM used, both single-variable and composite [22 – 26]. Although such a plurality of measures may be useful for etiologic research, in the case of monitoring, such heterogeneity impedes comparing results across studies, across outcomes, regions, and over time.

The utility of linking public health data to US census-based socioeconomic data to assess socioeconomic inequalities in health was first recognized in the 1920s and 1930s, in pathbreaking studies supported by the National Tuberculosis Association, following establishment of the first census tracts in New York City in 1906. These investigations, listed below, assessed people's risk of TB and later other health outcomes in relationship to socioeconomic conditions of their census tracts, which initially were also termed "sanitary areas" because of their utility for public health planning.

- Nathan WB. Health conditions in North Harlem 1923-1927. New York: National Tuberculosis Association, 1932.
- Green HW. Tuberculosis and economic strata, Cleveland's Five-City Area, 1928-1931. Cleveland, OH: Anti-Tuberculosis League, 1932.
- Green HW. The use of census tracts in analyzing the population of a metropolitan community. J Am Stat Assoc 1933; 28:147-153.
- Terris M. Relation of economic status to tuberculosis mortality by age and sex. Am J Public Health 1948; 38:1061-70.

For additional discussion of early use of census tract data in public health analyses, see:

- Watkins RJ. Introduction. In: Watkins RJ, Swift AL Jr, Green HW, Eckler AR. Golden Anniversary of Census Tracts, 1956. Washington, DC: American Statistical Association; 1956:1-2.
- Coulter EJ, Guralnick L. Analysis of vital statistics by census tract. J Am Stat Assoc 1959;54:730-40.

We accordingly launched the Public Health Disparities Geocoding Project to ascertain which ABSMs, at which geographic level (census block group [BG], census tract [CT], or ZIP Code [ZC]), would be most apt for monitoring US socioeconomic inequalities in the health. To provide a robust evaluation, guided by ecosocial theory [32, 33], we designed the study to encompass a wide variety of health outcomes, hypothesizing that some ABSMs and geographic levels might be more sensitive to socioeconomic gradients for some health outcomes than others. Drawing on 1990 census data and public health surveillance systems of 2 New England states, Massachusetts and Rhode Island, we included 7 types of outcomes: mortality (all cause and cause-specific), cancer incidence (all-sites and site-specific), low birth weight, childhood lead poisoning, sexually transmitted infections, tuberculosis, and non-fatal weapons-related injuries [31, 34 – 37].

We likewise hypothesized that some socioeconomic measures might be more sensitive than others to socioeconomic gradients in health, and so analyzed socioeconomic gradients in relation to 18 ABSMs: 11 single-variable and 7 composite (**Table 1: Area-based socioeconomic measures: constructs and operational definitions, using 1990 US census data**). Pertinent a priori considerations to decide which measure(s) at which geographic level(s) would be best suited for monitoring socioeconomic gradients in health across diverse outcomes and within diverse racial/ethnic-gender groups were derived in part from Rossi and Gilmartin's criteria for valid and useful social indicators [38], and included: (a) external validity (do the measures find gradients in the direction reported in the literature, i.e., positive, negative, or none, and across the full range of the distribution?), (b) robustness (do the measures detect expected gradients across a wide range of outcomes?),

(c) completeness (is the measure relatively unaffected by missing data?), and (d) user-friendliness (how easy is the measure to understand and explain?).

Based on our methodologic research (see our appended published papers), our key methodologic finding was that the ABSM most apt for monitoring socioeconomic inequalities in health was the census tract (CT) poverty level [35 – 37, 39 – 40]. Specifically, we demonstrated that **the CT poverty measure:**

- consistently detected expected socioeconomic gradients in health across a wide range of health outcomes, among both the total population and diverse racial/ethnic-gender groups;
- yielded maximal geocoding and linkage to area-based socioeconomic data (compared to BG and ZC data), and
- was readily interpretable to and could feasibly be used by state health department staff

Indeed, fully 98% of our records could be geocoded to CT level, and data on poverty was missing for only 0.7% of the catchment area's CTs. We also demonstrated that:

- accuracy of geocoding, not just completeness, matters [39];
- ZIP Code data should not be used, because of biases introduced by the spatiotemporal mismatch of ZIP Code and US Census data [40]; and
- some socioeconomic measures (e.g., pertaining to wealth and to income inequality) were particularly insensitive to the expected socioeconomic gradients observed with the poverty measure and other ABSMs designed to measure economic deprivation.

**Based on these considerations, we arrived at our recommendation that the CT level measure of "percent of persons below poverty" would be most apt for monitoring US socioeconomic inequalities in health.**

In **Figure 2**, we show what socioeconomic gradients in health would look like, across our varied outcomes, if routinely monitored using the CT poverty measure. According to the US Census Bureau, CTs are "small, relatively permanent statistical subdivision of a county … designed to be relatively homogeneous with respect to population characteristics, economic status, and living conditions" and on average contain 4,000 persons [41, pp.G – 10, G – 11]. For 1990 census data, the poverty line (which varies by household size and age composition) equaled $12,647 for a family of 2 adults and 2 children [42]. In this Figure, we employ the following a priori cut points for the CT measure "percent of persons below poverty," based on our prior analyses [34 – 37]: 0-4.9%, 5.0-9.9%, 10.0-19.9%, and >=20%, the federal definition of a "poverty area. [43]"

Using this measure, we were able to provide evidence of powerful socioeconomic gradients not only for mortality and low birthweight, as has been well documented [1, 4, 9, 10], but also for myriad other outcomes for which socioeconomic data in the US are not routinely available: sexually transmitted infections, tuberculosis, violence, cancer incidence, and childhood lead poisoning. Additionally advantages were that:

- We were able to assess socioeconomic gradients in health, within the total population and diverse racial/ethnic-gender groups using a consistent socioeconomic measure across all outcomes, from birth to death, thereby avoiding well-known problems with individual-level measures of education and occupation (e.g., how to classify children and others who have not completed their education or who are not in the paid labor force) [22].
- We could show that adjusting solely for CT poverty substantially reduced excess risk observed in the black and Hispanic compared to white population.
- We likewise could generate what to our knowledge is the first statewide data on the population attributable fraction in relation to poverty, whereby we found that for half the outcomes over 50% of cases overall would have been averted if everyone's risk equaled that of persons in the least

impoverished CT, the only group that consistently achieved Healthy People 2000 goals a decade ahead of time.
- Lastly, the approach we employed permitted documenting the temporal persistence—and worsening status of—a previously identified "zone of excess mortality."

Equally salient, our method relied solely on appending nationally-available and widely-accessible US census data to the relevant public health records, thereby generating state-level data that could be aggregated up to national-level data, to monitor national trends in socioeconomic inequalities in health. Indeed, a recently issued monograph from the National Cancer Institute, on Area Socioeconomic Variations in U.S. Cancer [44], does just this: following the recommendation of our project, it employed the census-derived poverty measure at the tract level, where feasible, or otherwise at the county level, to document socioeconomic inequalities in cancer incidence, stage, treatment, survival, and mortality.

Importantly, the methodology we employed does not treat CT-level measures as a "proxy" for individual-level measures. Rather, it posits that ABSMs capture a mix of individual- and/or area-based socioeconomic effects, if extant. Likely at issue are a complex combination of 3 factors: (1) composition (people in poor areas have poor health because poor people, as individuals, have poor health), (2) context (people in poor areas also have poor health because concentration of poverty creates or exacerbates harmful social interactions), and (3) location of public goods or environmental conditions (poor areas are less likely to have good supermarkets and are more likely to be situated next to industrial plants, thereby harming health of their residents) [45, 46]. Were the relevant data available, these complex interactions could be analyzed using multilevel methods [45 – 47]. Even absent these more detailed data, however, using only ABSMs we could still detect marked—yet typically undocumented–socioeconomic gradients in health within diverse racial/ethnic-gender groups plus provide conservative estimates of their contribution to racial/ethnic health disparities.

Even so, caution is required regarding interpretation of our data in relation to race/ethnicity. This is because our estimates of the magnitude of socioeconomic inequalities in health, within and across diverse racial/ethnic groups, necessarily are subject to concerns about racial/ethnic misclassification and the census undercount [4, 9, 10]. By itself, the method of geocoding and employing area-based socioeconomic measures cannot directly address these two problems, which affect all population-based analyses reliant on public health surveillance and census data [44, 48, 49]. Recent analyses, however, suggest that these problems result in estimates of US death rates among the white and black population being overstated in official publications by only 1% and 5%, respectively, and being understated, by a similar degree, for Hispanics (by 2%), but by a much larger degree for American Indians (by 21%) and Asian or Pacific Islanders (by 11%) [48]. Similar patterns have been reported for cancer registry data [44, 50]and likely would affect the other outcomes (i.e., STI, TB, and injuries) also reliant on census denominators and total or partial use of non-self-report data on race/ethnicity. Such errors would result in a tendency to overestimate, compared to the white population, an excess risk among the black population and a reduced risk among the Hispanic population. Analyses of low birth weight and childhood lead poisoning, by contrast, would not be affected by the census undercount, since the denominators were, respectively, the births themselves and the children screened; moreover, racial/ethnic misclassification was minimized by use of self-report racial/ethnic data in these surveillance systems.

An additional caveat pertains to our use of the US poverty line as an indicator of socioeconomic deprivation. Although debates exist over how best to measure poverty in the US [51, 52], precisely because of its significance for policies and for resource allocation [51, 52], evidence indicates the CT poverty measure, especially in excess of 20% (the federal definition of a "poverty area" [43], does provide a reasonable decennial indicator of neighborhood economic deprivation, as assessed in relation to housing deterioration, refuse, crime, and other social indicators (e.g., unemployment, low earnings, low education) [43, 44, 52, 54]. Also underscoring the robustness of the CT poverty measure as a useful economic indicator, we found similar results [34 – 37]in analyses utilizing data on the percent of persons below 50% of the US poverty line, above 200% of the US poverty line,

and below 50% of the US median household income (an alternative measure of poverty employed in many European countries [53]. In all of these analyses, the magnitude of the socioeconomic gradients detected were on par with available estimates reported in the US [1, 4, 9, 10]and analogous European literature [2, 3, 25]., The net implication is use of the CT poverty measure is unlikely to overestimate either the extent of socioeconomic gradients or their contribution to racial/ethnic disparities in health, and instead provides a useful metric that reveals the widespread and often profound extent to which socioeconomic deprivation adversely shapes population health, from infancy to death.

In conclusion, results of our study highlight the importance—and feasibility–of routinely monitoring US socioeconomic inequalities in health, overall and stratified by race/ethnicity and gender, thereby painting a truer picture of the "public-health situation as a whole," as long urged by Sydenstricker and other public health leaders [14, 17, 20, 21, 55]. Addressing gaps in policy-relevant knowledge [1 – 3, 7 – 14, 16 – 18, 56, 57], the evidence generated by our approach could be used to set health objectives, guide resource allocation, and track progress—and setbacks–in reducing social disparities in both health and health care, at the national, state, and local level. Relying on widely-available data, the proposed methodology not only is cost-efficient but also permits comparisons within and across health outcomes throughout the US, over time, based on a common metric for socioeconomic position derived from US census data. Timeliness of CT data, moreover, will be improved, starting in 2008, when the American Community Survey starts releasing annual CT estimates, based on 5-year rolling averages [58]. Were data on US socioeconomic inequalities in health readily available, and reported upon yearly, for both the total population and diverse racial/ethnic-gender groups, efforts to track—and improve accountability for addressing—social disparities in health would be greatly enhanced. We suggest this can be accomplished by geocoding US public health surveillance data and using the CT-level measure "percent of persons below poverty."

In the rest of this monograph, we explain our methods to facilitate their use by others. Specific sections focus on:

- how we geocoded our data;
- how we constructed the ABSMs;
- how we tested these measures across diverse health outcomes at different geographic levels;
- how we generated our figures; and
- a guided exercise, using a sample data file, to facilitate trying out our approach, with steps clearly delineated and answers provided to check accuracy of implementation.

We hope you will find this monograph useful in improving efforts to monitor socioeconomic inequalities in health, both within the total population and diverse racial/ethnic-gender groups, thereby making a vital contribution to identifying and galvanizing action to address social disparities in health.

---

# REFERENCES

1. National Center for Health Statistics. Health, United States, 1998 with Socioeconomic Status and Health Chartbook. Hyattsville, MD: US Dept of Health and Human Services, 1998.

2. Evans T, Whitehead M, Diderichsen F, Bhuiya A, Wirth M (eds). Challenging inequities in health: from ethics to action. Oxford, UK: Oxford University Press, 2001.

3. Shaw M, Dorling D, Gordon D, Davey Smith G. The Widening Gap: Health Inequalities and Policy In Britain. Bristol, UK: The Policy Press, 1999.

4. Krieger N, Rowley DL, Herman AA, Avery B, Phillips MT. Racism, sexism, and social class: implications for studies of health, disease, and well-being. Am J Prev Med 1993; 9 (Suppl):82-122. https://pubmed.ncbi.nlm.nih.gov/35597564/

5. US Department of Health and Human Services. Healthy People 2010 (Conference edition, in two volumes). Washington, DC: US Govt Printing Office, 2000; Objective 23-3.

6. Krieger N, Chen JT, Ebel G. Can we monitor socioeconomic inequalities in health? A survey of U.S. Health Departments' data collection and reporting practices. Public Health Rep 1997; 112:481-91. https://pubmed.ncbi.nlm.nih.gov/10822475/

7. Friedman DJ, Hunter EL, Parrish RG. Shaping a Vision of Health Statistics for the 21st Century. Washington, DC: Department of Health and Human Services Data Council, Centers for Disease Control and Prevention, National Center for Health Statistics, and National Committee on Vital and Health Statistics, 2002.

8. Krieger N. The making of public health data: paradigms, politics, and policy. J Public Health Policy 1992; 13:412-427. https://pubmed.ncbi.nlm.nih.gov/1287038/

9. Williams DR, Collins C. US socioeconomic and racial differences in health: patterns and explanations. Annu Rev Sociol 1995; 21:349-86.

10. Kington RS, Nickens HW. Racial and ethnic differences in health: recent trends, current patterns, future directions. In: National Research Council. America becoming: racial trends and their consequences. Vol 2. Smelser NJ, Wilson WJ, Mitchell F (eds). Washington, DC: National Academy Press, 2001; 253-310.

11. Navarro V. Race or class versus race and class. Lancet 1990; 336:1238-40.

12. Pantazis C, Gordon D (eds). Tackling Inequalities: Where Are We Now and What Can Be Done? Bristol, UK: The Policy Press, 2000.

13. Marmot M, Wilkinson RG (eds). Social Determinants of Health. Oxford: Oxford University Press, 1999.

14. Porter D (ed). The History of Public Health and the Modern State. Amsterdam; Atlanta, GA: Rodopi, 1994.

15. National Center for Health Statistics. Health, United States 2002 with Chartbook on Trends in the Health of Americans. Hyattsville, MD: National Center for Health Statistics, 2002.

16. Braveman P, Starfield B, Geiger HJ. World Health Report 2000: how it removes equity from the agenda for public health monitoring and policy. Br Med J 2001;323:678-81.

17. Sydenstricker E. The incidence of illness in a general population group: General results of a morbidity study from December 1, 1921 through March 31, 1924, Hagerstown, Md. Public Health Rep 1925;40:279-91.

18. Sydenstricker E. Health and the Depression. Milbank Memorial Fund Quarterly 1934: 12:273-280.

19. US Department of Health, Education, and Welfare. Origin and Program of the U.S. National Health Survey. Health Statistics Series A1, May 1958, p. 3.

20. Winslow C-EA. The Life of Hermann M. Biggs, M.D., D.Sc., LL.D, Physician and Statesman of the Public Health. Philadelphia, PA: Lea & Febiger, 1929.

21. Rosen G. A History of Public Health. (1958). Introduction by Elizabeth Fee; Bibliographical essay and new bibliography by Edward T. Morman. Expanded ed. Baltimore, MD: Johns Hopkins University Press, 1993.

22. 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. https://pubmed.ncbi.nlm.nih.gov/9143723/

23. 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. https://pubmed.ncbi.nlm.nih.gov/1566949/

24. Lynch J, Kaplan G. Socioeconomic position. In: Berkman L, Kawachi I (eds). Social Epidemiology. Oxford: Oxford University Press, 2000; 13-35.

25. Lee P, Murie A, Gordon D. Area Measures of Deprivation: A Study of Current Methods and Best Practices in the Identification of Poor Areas in Great Britain. Birmingham, UK: Centre for Urban and Regional Studies, University of Birmingham, 1995.

26. Carstairs V. Socio-economic factors at areal level and their relationship with health. In: Elliott P, Wakefield J, Best N, Briggs D (eds). Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press, 2000; 51-67.

27. Nathan WB. Health Conditions In North Harlem 1923-1927. New York: National Tuberculosis Association; 1932.

28. Green HW. The use of census tracts in analyzing the population of a metropolitan community. J Am Stat Assoc 1933; 28:147-153.

29. Terris M. Relation of economic status to tuberculosis mortality by age and sex. Am J Public Health 1948; 38:1061-70.

30. Coulter EJ, Guralnick L. Analysis of vital statistics by census tract. J Am Stat Assoc 1959; 54:730-40. back to top

31. Krieger N, Zierler S, Hogan JW, Waterman P, Chen J, Lemieux K, Gjelsvik A. Geocoding and measurement of neighborhood socioeconomic position. In: Kawachi I, Berkman LF (eds). Neighborhoods and Health. New York: Oxford University Press, 2003; 147-178. https://academic.oup.com/book/6120

32. Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Soc Sci Med 1994; 39:887-903. https://pubmed.ncbi.nlm.nih.gov/7992123/

33. Krieger N. Theories for social epidemiology for the 21st century: an ecosocial perspective. Int J Epidemiol 2001; 30:668-677. https://pubmed.ncbi.nlm.nih.gov/11511581/

34. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. 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 2002; 156:471-482. https://pubmed.ncbi.nlm.nih.gov/12196317/

35. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). J Epidemiol Community Health 2003; 57:186-199. https://pubmed.ncbi.nlm.nih.gov/12594195/

36. Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian S. Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures–The Public Health Disparities Geocoding Project (US). Public Health Rep 2003; 118:240-260. https://pubmed.ncbi.nlm.nih.gov/12766219/

37. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. 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 2003; 93: 1655-1671. https://pubmed.ncbi.nlm.nih.gov/14534218/

38. Rossi RJ, Gilmartin KJ. The Handbook of Social Indicators: Sources, Characteristics, and Analysis. New York, NY: Garland STPM Press, 1980.

39. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001; 91:1114-1116. https://pubmed.ncbi.nlm.nih.gov/11441740/

40. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–The Public Health Disparities Geocoding Project. Am J Public Health 2002; 92:1100-1102. https://pubmed.ncbi.nlm.nih.gov/12084688/

41. US Bureau of the Census. Geographical Areas Reference Manual. Washington, DC: US Dept of Commerce, 1994.

42. US Bureau of Census. Census of population and housing, 1990: Summary Tape File 3 technical documentation. Washington, DC: Bureau of the Census, 1991.

43. US Bureau of the Census. Poverty areas.

44. Singh GK, Miller BA, Hankey BF, Edwards BK. Area Socioeconomic Variations in U.S. Cancer Incidence, Mortality, Stage, Treatment, and Survival, 1975-1999. NCI Cancer Surveillance Monograph Series, Number 4. Bethesda, MD: National Cancer Institute, 2003. (NIH Pub. No. 03-5417).

45. Macintyre S, Ellaway A, Cummins S. Place effects on health: how can we conceptualise, operationalise and measure them? Soc Sci Med 2002;55:125-39.

46. O'Campo P. Invited commentary: advancing theory and methods for multilevel models of residential neighborhoods and health. Am J Epidemiol 2003;157:9-13.

47. Subramanian SV, Jones K, Duncan C. Multilevel methods for public health research. In: Kawachi I, Berkman L (eds). Neighborhoods and Health. Oxford: Oxford University Press 2003:65-111.

48. Rosenberg HM, Maurer JD, Sorlie PD, Johnson NJ, MacDorman MF, Hoyert DL, Spitler JF, Scott C. Quality of death rates by race and Hispanic origin: a summary of current research, 1999. National Center for Health Statistics, Vital Health Stat 2(128), 1999.

49. Keppel KG, Pearcy JN, Wagener DK. Trends in racial and ethnic-specific rates for health status indicators: United States, 1990-1998. Healthy People 2000 Statistical Notes, no. 23. Hyattsville, MD: National Center for Health Statistics, January 2002.

50. United States Cancer Statistics Working Group. United States Cancer Statistics: 1999 Incidence. Department of Health and Human Services, Atlanta, GA: Centers for Disease Control and Prevention and National Cancer Institute 2002.

51. Citro CF, Michael RT (eds). Measuring Poverty: A New Approach. Panel on Poverty and Family Assistance: Concepts, Information Needs, and Measurement Methods. Washington, DC: National Academy Press, 1995.

52. Citro CF, Kalton G (eds). Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press, 2000.

53. Gordon D, Spicker P (ed). The International Glossary on Poverty. London: Zed Books, 1999.

54. Jargowsky PA. Poverty and Place: Ghettos, Barrios, and The American City. New York: Russell Sage, 1997.

55. Krieger N, Fee E. Measuring social inequalities in health in the United States: an historical review, 1900-1950. Int J Health Serv 1996;26:391-418. https://pubmed.ncbi.nlm.nih.gov/8840195/

56. Fiscella K, Franks P, Gold M, Clancy C. Inequalities in quality: addressing socioeconomic, racial and ethnic disparities in health care. JAMA 2000; 283:2579-84.

57. Friedman D, Anderka M, Krieger J, Land G, Solet D. Accessing population health information through interactive systems: lessons learned and future directions. Public Health Rep 2001; 116:132-41.

58. US Census Bureau. Survey Basics: What is the American Community Survey? (Accessed on February 5, 2004 at: http://www.census.gov/acs/www/SBasics/What/What1.htm ).

# Publications

Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: The Public Health Disparities Geocoding Project. Am J Public Health 2005; 95: 312-323. https://pubmed.ncbi.nlm.nih.gov/15671470/

Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. 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 2003; 93: 1655-1671. https://pubmed.ncbi.nlm.nih.gov/14534218/

Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian S. Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures–The Public Health Disparities Geocoding Project (US). Public Health Rep 2003; 118:240-260. https://pubmed.ncbi.nlm.nih.gov/12766219/

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). J Epidemiol Community Health 2003; 57:186-199. https://pubmed.ncbi.nlm.nih.gov/12594195/

Subramanian SV, Chen JT, Rehkopf DH, Waterman PD, Krieger N. Comparing Individual- and Area-based Socioeconomic Measures for the Surveillance of Health Disparities: A Multilevel Analysis of Massachusetts Births, 1989-1991. Am J Epidemiol 2006; 164:823-834, https://pubmed.ncbi.nlm.nih.gov/16968866/

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. 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 2002; 156:471-482. https://pubmed.ncbi.nlm.nih.gov/12196317/

Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–The Public Health Disparities Geocoding Project. Am J Public Health 2002; 92:1100-1102. https://pubmed.ncbi.nlm.nih.gov/12084688/

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001; 91:1114-1116. https://pubmed.ncbi.nlm.nih.gov/11441740/

Krieger N, Zierler S, Hogan JW, Waterman P, Chen J, Lemieux K, Gjelsvik A. Geocoding and measurement of neighborhood socioeconomic position. In: Kawachi I, Berkman LF (eds). Neighborhoods and Health. New York: Oxford University Press, 2003; 147-178. https://academic.oup.com/book/6120

Subramanian SV, Chen JT, Rehkopf DH, Waterman PD, and Krieger N. Racial Disparities in Context: A Multilevel Analysis of Neighborhood Variations in Poverty and Excess Mortality Among Black Populations in Massachusetts. Am J Public Health 2005; 95: 260-265. https://pubmed.ncbi.nlm.nih.gov/15671462/

Krieger N. A century of census tracts: health and the body politic (1906 – 2006) J Urban Health ("in press" at time monograph was published; actual publication: 2006; 83(3):355-361). https://pubmed.ncbi.nlm.nih.gov/16739037/

Chen JT, Rehkopf DH, Waterman PD, Subramanian SV, Coull BA, Cohen B, Ostrem M, Krieger N. Mapping and measuring social disparities in premature mortality: the impact of census tract poverty within and across Boston neighborhoods, 1999-2001. J Urban Health 2006; 83(6):1063-1084. https://pubmed.ncbi.nlm.nih.gov/17001522/

Subramanian SV, Chen JT, Rehkopf DH, Waterman PD, Krieger N. Subramanian et al. Respond to "Think Conceptually, Act Cautiously." Am J Epidemiol 2006; 164:841-844. https://doi.org/10.1093/aje/kwj315

Rehkopf DH, Haughton LT, Chen JT, Waterman PD, Subramanian SV, Krieger N. Monitoring Socioeconomic Disparities in Death: Comparing Individual-Level Education and Area-Based Socioeconomic Measures. Am J Public Health 2006; 96: 2135-2138 (with erratum in AJPH 2007; 97:1543) https://pubmed.ncbi.nlm.nih.gov/16809582/

Krieger N. Why Epidemiologists Cannot Afford to Ignore Poverty.[Editorial] Epidemiology 2007; 18:658-663. https://pubmed.ncbi.nlm.nih.gov/18049180/

Krieger N. Putting health inequities on the map: social epidemiology meets medical/health geography — an ecosocial perspective. Geojournal 2009; 74:87-97. https://link.springer.com/article/10.1007/s10708-009-9265-x

Krieger N, Waterman PD, Chen JT, Subramanian SV, Rehkopf DH. Monitoring socioeconomic determinants for healthcare disparities: tools from the Public Health Disparities Geocoding Project. In: Williams RA (eds). Eliminating Healthcare Disparities in America: Beyond the IOM Report. Totowa, NJ: Humana Press, 2007; 259-306. https://link.springer.com/book/10.1007/978-1-59745-485-8

# How To…

- We provide a primer on the basics of **Geocoding**, including descriptions of the many options and services available, and the nitty-gritty details of address cleaning, address formatting, and evaluation of geocoding accuracy.
- In **Generating ABSMs** we describe the concepts, methods, and measures behind creating area-based socioeconomic measures, including a summary table of the 19 theoretically justified area-based socioeconomic measures we created based on 1990 U.S. Census data (see our **ABSM Creation Table**).
- Under **Analytic Methods**, we provide details on how to merge geocoded surveillance data with Census derived population denominators and area-based socioeconomic measures. We also present basic epidemiologic methods for generating descriptive statistics, including directly age-standardized incidence rates, incidence rate ratios and rate differences, the relative index of inequality, and population attributable fraction. Examples are provided for each of these techniques, and each section is further linked to a comprehensive **Case Example**.
- We've also included some information about **Multi-level Modeling** and **Visual Display** of data for surveillance reporting.

# Geocoding

## Geocoding vs. GIS

*GIS and Geocoding are two terms that you've probably been hearing a lot about recently. What are they exactly?*

**GIS** – Geographical Information Systems – are technology based systems that combine layers of geographic data to give you a better understanding of a particular place [1]. For example, you might combine a layer of cholera outbreaks with a layer of water sources to be able to display graphically the relationship between the two. For more examples of GIS technology at work, visit www.esri.com.

**Geocoding** is the assignment of a code – usually numeric — to a geographic location. (So, one geocode that you're probably already familiar with is your ZIPcode.) Usually however, when someone talks about geocoding, they are talking about geocodes that are a bit more specific, i.e., affixing to an individual address its latitude and longitude – which is, very simply, the vertical and horizontal distance of a point relative to the equator [2]. Once the latitude and longitude are known, you can then figure out all sorts of other geocodes to affix by determining what geographic regions the specified point lies in, e.g., what ZIPcode does this point lie in? What census tract? What census blockgroup? What police precinct? Appending any of these codes to a specific street address is considered geocoding.

Our project utilized primarily geocoding technology. Before continuing to discuss geocoding, it's important that you know a bit about census geography, since the geographic code that is typically affixed to an address during geocoding is either the U.S. Census Bureau defined census tract or blockgroup.

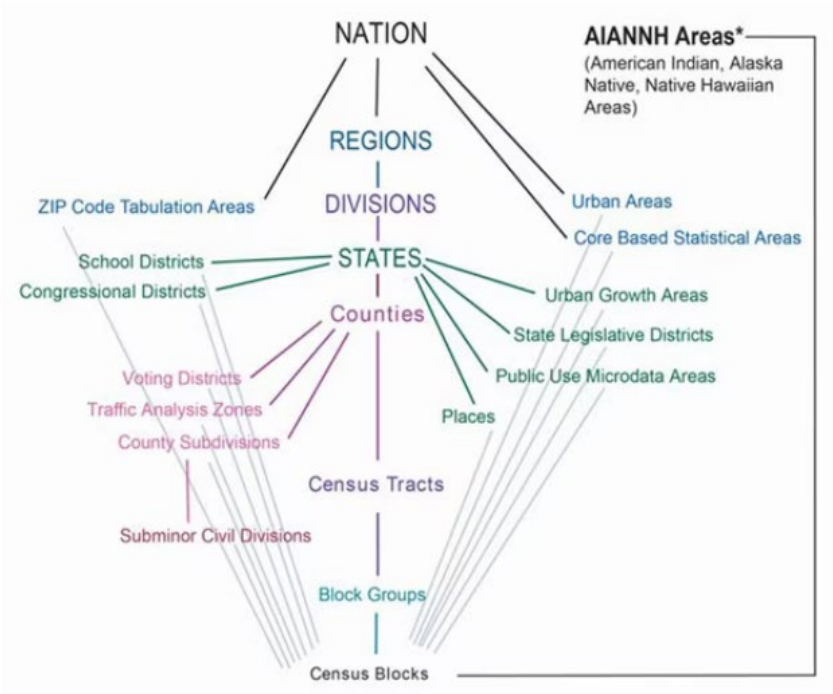


Figure 2. Geographic Hierarchy for the 1990 decennial Census

The above figure displays the hierarchy of census geography [3]. As you may have already read in other sections of this monograph, we strongly recommend the census tract as the geographical unit of analyses. Census tracts, census blockgroups, and the "new to 2000" ZIP Code Tabulation Areas are U.S. Census Bureau defined, standardized, and relatively permanent geographical units. Census tracts are constructed specifically to include on average 4,000 people of fairly homogeneous population characteristics, economic position, and living conditions. Federal, state, and local governments routinely use census tracts as administrative units. For example, the Federal government uses census tracts to define urban empowerment zones and decide who's eligible for low-income housing tax credits. Census tracts are sub-divided into blockgroups — which have an average population size of about 1,000.

See Figure 2 below. Notice (in the figure above) that ZIPcodes are off to the side, in a category all by themselves, and not linkable to anything else. In contrast to the census tracts and blockgroups, ZIP codes are U.S. Postal Service administrative units that are subject to change at any time, thus making the linking of ZIPcode level data to other datasets, e.g., the decennial U.S. Census data a bit questionable. They are far from standardized – a ZIPcode can designate a single office building or entire state county. For a more thorough discussion of the problem of using ZIPcodes in area-based analyses, please refer to our article "Zip Code caveat: bias due to spatiotemporal mismatches between ZIP Codes and US census-defined areas—the Public Health Disparities Geocoding Project" [4].

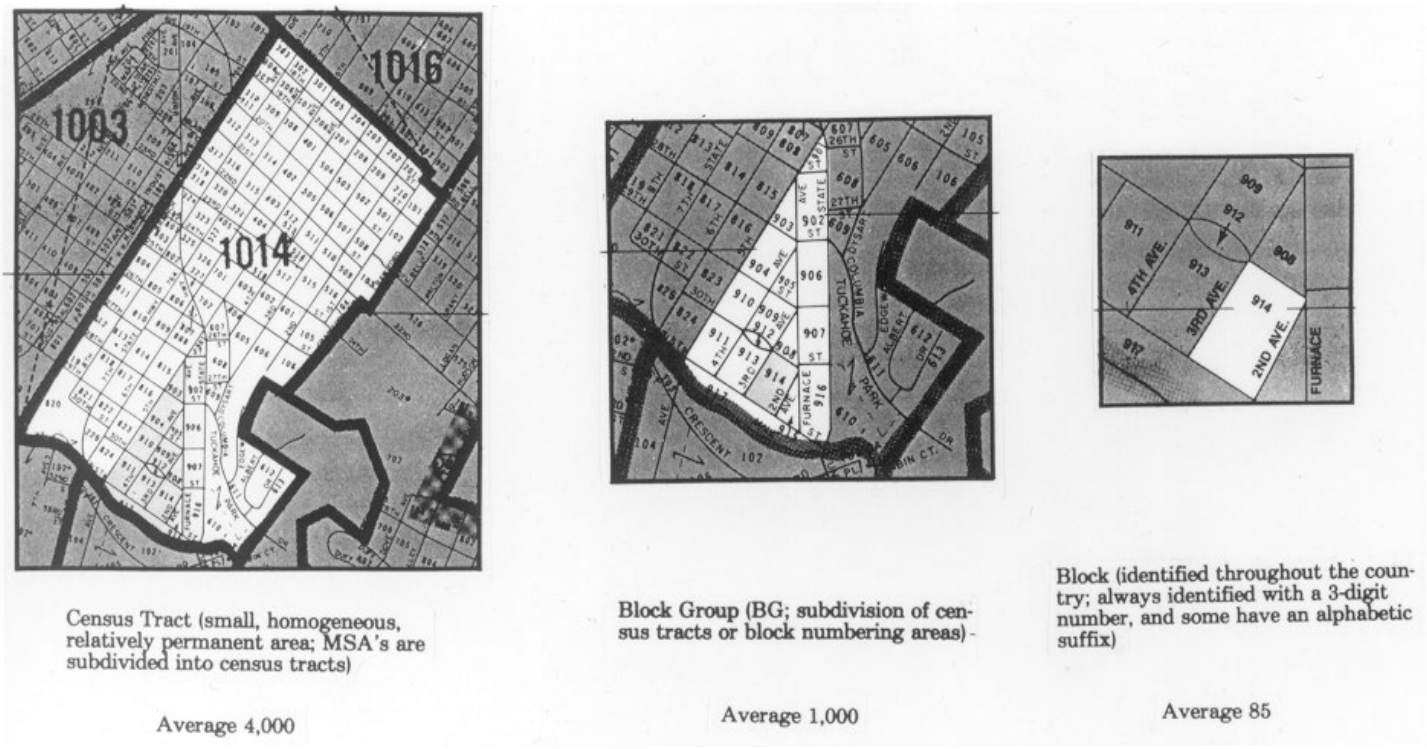



Figure 3. Census Tracts, Blockgroups, and Blocks

For this project, we geocoded Massachusetts Department of Public Health and Rhode Island Department of Health data to the blockgroup level. Before we eventually used a commercial geocoding firm to geocode our data, we considered three things: accuracy, cost, and turnaround time. In 1999, when we explored our geocoding options, there were a handful of commercial services and two stand-alone geocoding programs available. Now, there are many more commercial geocoding services, and quite a few stand-alone programs to

choose from. However, not all companies or programs are the same. So, first determine what makes the most sense for your project: using a geocoding service or using a program to do the geocoding yourself.

Considerations include time – are you working on a tight schedule, or do you have enough time for you, or someone on your staff, to become proficient with a geocoding program? Keep in mind that some of the programs have very steep learning curves. Becoming proficient at geocoding will take months, and becoming an expert may take years. The benefit to having a trained in-house geocoding specialist is that, over time, depending on the volume of your data, it may be cheaper to geocode in-house, and you have the additional benefit of having more control over the geocoding process.

If you decide to use a geocoding service, we recommend that you do a bit of testing to make sure you get the most accurate results. Many companies advertise high completion rates, that is, the percentage of addresses that they geocode, but completeness and accuracy are two different things. How do you know if they've geocoded the addresses to the right place? To test the accuracy of geocoding services, we recommend the following plan6.

First, generate a test file. This you'll do by performing some "old-fashioned" geocoding. Pull together a list of 50-75 addresses that you're familiar with. They should be spread across as large a geographic area as possible, but concentrated in the area that the majority of your data (the addresses you are eventually planning to geocode) will be from. On a street map (or more than one street map if your addresses cover a large enough area), locate and mark the exact locations of the addresses. Take this map to your regional Census Bureau office. Using the official Census Bureau blockgroup maps available there, identify the blockgroup that each address falls in. To create the full blockgroup geocode, use the following scheme:

Digits 1-2 = State code
Digits 3-5 = County code
Digits 6-11 = Census Tract code (often used with a decimal point:xxxx.xx)
Digit 12 = Blockgroup code


You'll be able to get all of the components that make up the areakey from the blockgroup maps at the census bureau. For more information about blockgroups and other units of census geography, check out The Census Geographic Areas Reference Manual [3]. (The U.S. Census Bureau Website is a great place to familiarize yourself with a lot of subjects that we'll be focusing on in this monograph, e.g., Census data, Census geography, area-based measures, geocoding, GIS, and mapping.)

Congratulations! You've just (a) successfully geocoded your data to the blockgroup level; and (b) created a test file. You can now use this file to test commercial geocoding firms and geocoding programs alike.

Send a file containing only the addresses to the prospective geocoding companies and then compare the results sent back from the company to the correct geocodes you ascertained at the Census Bureau office. As an external check of both you and the geocoding companies, submit your addresses to the Census Bureau Census Tract Locator on the American FactFinder website. If you opted to use a geocoding program, you can also use this test file to test your own results.

Now that you have your geocoding plan of action ready – whether it's using a geocoding program yourself or sending your data out to a geocoding service — the next step is to clean your addresses. Geocoding follows the

time tested theorem “garbage in, garbage out”. If your addresses are not clean, then you are significantly increasing the probability that they will not be geocoded correctly.

Cleaning addresses means:

- retaining only the key address elements in one field: house/building number; street name; street type; e.g., 100 Main St
- getting rid of all extraneous characters, e.g., “BSMT” “REAR” “APT 1” “UNIT 3”, etc.
- standardizing spelling, e.g., converting all incidences of “Mass Ave” to “Massachusetts Ave”

Some examples:

| Record # | Original Address | “Cleaned” |
|---|---|---|
| 1 | 677 Huntington, #304 | 677 Huntington Ave |
| 2 | 46 Burr REAR | 46 Burr St |
| 3 | Unit B, 1200 Comm Ave. | 1200 Commonwealth Ave |
| 4 | 423 Allston St., 4th Floor, Suite 100 | 423 Allston St |
| 5 | The Landmark Building, 401 Park Drive | 401 Park Drive |
| 6 | 99 ½ Chauncey St | 99 Chauncy St |

What about those pesky P.O. Box addresses and “Rural Route” addresses with no house numbers?

• The geocoding program will look at “P.O. Box” as if it’s a street name, so if there’s a “Postbox St.” in your neighborhood, you may get false matches.

• The individual who has this P.O.Box as a mailing address may not necessarily live in the blockgroup, census tract, or even the ZIPcode that the post office is in.

• Check a map. Does the entire rural route lie in a single census tract? Or in a single blockgroup? If so, the geocodes may be accurate since ALL structures on that route fall in the same census tract.

• Decide ahead of time on a method of dealing with P.O. Boxes and Rural Route addresses in your analyses. Keep in mind that these addresses are often not geocodable anyway.

For more detail about cleaning and formatting addresses, speak with the Customer Service representative at the geocoding service, or check to see what format your geocoding program requires. Also note that there are a number of products on the market that will clean addresses for you. We have not evaluated them however, and so can not advise you regarding their efficacy or accuracy.

The typical format of a file to be sent to a geocoding service
(Excel or dbf format):

| Record # | Street Address | City | State | ZIPcode |
|---|---|---|---|---|
| 1 | 677 Huntington Ave | Boston | MA | 2115 |
| 2 | 46 Burr St | Jamaica Plain | MA | 2130 |
| 3 | 1200 Commonwealth Ave | Boston | MA | 2215 |

| | | | | |
|---|---|---|---|---|
| 4 | 423 Allston St | Cambridge | MA | 2139 |
| 5 | 401 Park Drive | Boston | MA | 2215 |
| 6 | 99 Chauncy Street | Boston | MA | 2111 |

The typical format of a file returned from a geocoding service (Excel or dbf format):

| Record # | Street Address | City | State | ZIPcode | Latitude | Longitude | Areakey | Match Code |
|---|---|---|---|---|---|---|---|---|
| 1 | 677 Huntington Ave | Boston | MA | 2115 | -71.1 | 42.34 | 25025081000 | AS0 |
| 2 | 46 Burr St | Jamaica Plain | MA | 2130 | -71.11 | 42.32 | 25025120600 | AS1 |
| 3 | 1200 Commonwealth Ave | Boston | MA | 2215 | -71.12 | 42.35 | 25025000801 | AS7 |
| 4 | 423 Allston St | Cambridge | MA | 2139 | -71.11 | 42.36 | 25017353200 | ZB7I |
| 5 | 401 Park Drive | Boston | MA | 2215 | -71.1 | 42.34 | 25025010200 | AS0 |
| 6 | 99 Chauncy Street | Boston | MA | 2111 | -71.06 | 42.35 | 25025070100 | ZB7L |

The MatchCode variable (also called "georesult" by some companies) is an indicator of which address elements determined the geocode, and how certain the geocoding program is about the accuracy of the geocode. For example, the MatchCode of AS0 indicates that the geocode was derived based on the street address and matched exactly to a street segment in the program; the program is certain of blockgroup level accuracy. A MatchCode of ZC5Y indicates that the geocode assigned is based upon the location of the post office that delivers mail to that address, and the geocoding program is only comfortable claiming county level accuracy. (This is typically the MatchCode assigned to a P.O.Box address.)

A full explication of the MatchCodes will be provided to you by the geocoding service you employ, or in the technical notes of the program that you use.

Once you have your geocoded file, you should check for any discrepancies in the geocoding. Use SAS, or some other data analyses program, to look for differences in match rates by your variables of interest. At the very least, check for differences in geocoding rates by age, gender, race/ethnicity, socioeconomic data (if available). What are possible explanations for these differences – and how will they affect your analyses?

---

# REFERENCES


1. GIS.COM.

2. Stern, DP. From Stargazers to Starships.

3. Bureau of the Census, U.S. Department of Commerce. Geographic Areas Reference Manual. Washington, DC: Bureau of the Census, 1994.

4. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–The Public Health Disparities Geocoding Project. Am J Public Health 2002; 92:1100-1102. https://pubmed.ncbi.nlm.nih.gov/12084688/

5. U.S. Department of Commerce, Census ’90 Basics. Washington, D.C.: U.S. Government Printing Office, 1990.

6. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001; 91:1114-1116. https://pubmed.ncbi.nlm.nih.gov/11441740/

# Generating ABSMs


## Generating ABSMs: concepts, methods, and measures

Generating area-based measures of socioeconomic position requires an explicit approach to understanding what socioeconomic inequality is and how to measure it, at multiple levels. In this section we briefly review our definitions of "social class" and "socioeconomic position," and then delineate our approach to generating and appraising the validity and utility of our Project's area-based socioeconomic measures (ABSMs).

## Definitions: social class and socioeconomic position

Starting first with definitions, in the Public Health Disparities Geocoding Project we used the construct of "social class" to refer to social groups arising from interdependent economic relationships among people [1 – 2]. Stated simply, broad classes–like the working class, business owners, and their managerial class–exist in relationship to and co-define each other. One cannot, for example, be an employee if one does not have an employer and this distinction–between employee and employer–is not about whether one has more or less of a particular attribute, but concerns one's relationship to work and to others through a society's economic structure. Also at issue is an asymmetry of economic relations, whereby owners of resources (e.g., capital) gain economically from the labor or effort of employees.

Class, as such, is therefore logically and materially prior to its manifest expression in what can be referred to as socioeconomic position, an aggregate concept that includes both resource-based and prestige-based measures, as determined by both childhood and adult social class position [1]. Resource-based measures refer to material and social resources and assets, including income, wealth, and educational credentials; terms used to describe inadequate resources include "poverty" and "deprivation." Prestige-based measures refer to individuals' rank or

status in a social hierarchy. The term "socioeconomic status" should accordingly be avoided both because it arbitrarily (if not intentionally) privileges "status" over material resources as a determinant of health and because it conflates pathways involving material resources with those involving psychosocial appraisals of relative status.

# Measuring socioeconomic position: domains, levels, & lifecourse

Key domains of socioeconomic position relevant to understanding population health thus include [1 – 2]:

- Occupational class, which can affect health both directly and indirectly, via occupational hazards and via wages or income, relevant to standard of living;
- Educational attainment/credentials, usually reflective of childhood socioeconomic position and relevant to future economic prospects, and also relevant vis a vis knowledge & health literacy;
- Income & entitlements/subsidies, together affecting standard of living, noting that what "income" buys in a given society is related in part to what is provided by the social wage;
- Wealth, referring to accumulated assets, with important distinctions between what's readily fungible or not (e.g., stocks vs equity in a home), plus also wealth's converse, i.e., debt; and
- Relative social ranking, typically referring to "status" & "prestige."

Second, each domain can be assessed at multiple levels, including: individual, household, and area or neighborhood, plus also regional, national and global [1 – 5].

Likewise, relevant moments during the lifecourse for which one may want socioeconomic data include: in utero, infancy, childhood, plus early, middle, and late adulthood [1 – 2].

# From this vantage, we opted to create a variety of ABSMs, intended to capture diverse domains of SEP, for diverse outcomes that spanned the lifecourse, literally from birth until death.

# Formulating the ABSMs from census data

As described more fully in our Project's manuscripts [6 – 9], we created 19 theoretically-justified ABSMs (11 single variable, 8 composite ABSMS), delineated in the ABSM Creation Table. Two criteria central to formulating these ABSMS for socioeconomic position (SEP) were that they: (a) meaningfully summarized important aspects of the specified area's socioeconomic conditions, and (b) employed socioeconomic data that could legitimately be compared over time and across regions [1 – 9]. Based on our a priori conceptual definitions of SEP and social class1 and US, UK, and other global evidence emphasizing detrimental effects of material deprivation on health [10 – 14], we developed ABSMs for 6 domains of SEP: occupational class, income, poverty, wealth, education, and crowding, premised on the understanding that social class, as a social relationship, fundamentally drives the distribution of these manifest aspects of SEP [1]. Of note, one measure we included differs from the others: the Gini coefficient, which is a measure of within-area socioeconomic inequality rather than a measure of the average socioeconomic level of an area [15]. We included it because of concerns expressed about its uncritical use at the BG and CT level, given realities of economic segregation [9, 16].

Operationally, we generated each ABSM at each level of geography for each state. Among the composite variables, two were US analogues of the UK Townsend [4 – 5, 17] and Carstairs [4 – 5, 18] deprivation indices, one used the algorithm for the US Center for Disease Control and Prevention's "Index of Local Economic

Resources, [19]" and five were created exclusively for our study. To mirror the skewed population distribution of socioeconomic resources, "SEP1" and "SEP2" simultaneously combined categorical data on poverty, working class, and either wealth or high income. Finally, we produced an "SEP index" akin to the Townsend index, based on summation of standardized z scores of selected ABSM.

Lastly, our a priori criteria for evaluating the ABSMs pertained to: (1) external validity (did it detect the expected socioeconomic gradients for each outcome, i.e., positive, negative, or none at all?), (2) robustness (did the ABSM do so across multiple outcomes, as well as within diverse population groups?), (3) completeness (was the ABSM relatively unaffected by missing data?), and (4) user-friendliness (was it easy to understand and use?) [7 – 9, 22]. Based on these criteria, and the key findings we summarize in the introduction to this monograph, in the case example developed for this monograph, we focus solely on the census tract poverty ABSM.

# Example: Creating a single variable ABSM – % of persons below poverty

This section will describe how to create a single variable ABSM from census data, using the example of "percent of persons below poverty." As indicated in the ABSM creation table (**Area-based socioeconomic measures: constructs and operational definitions, using 1990 US census data**), the data for creating this variable for 1990 is available from census table P117, available from the summary tape file 3 (STF3). P117 gives population counts of persons above and below poverty, stratified by age. As an example, **p117** shows the counts for table P117 for all of Massachusetts.

To calculate the proportion of persons below poverty for this region, we sum all categories P1170001 to P1170024 to get the denominator, and sum categories P1170013 to P1170024 to get the numerator, and then simply divide this numerator by the denominator:

$$(P1170013 + \ldots + P1170024) / (P1170001 + \ldots + P1170024)$$

Creating a composite ABSM is similar in principle to creating a single variable ABSM, but requires a few extra steps. The Townsend ABSM consists of four components, percent crowding, percent unemployment, percent of individuals who do not own cars, and percent renters. The ABSM creation table (**Area-based socioeconomic measures: constructs and operational definitions, using 1990 US census data**) indicates where these four variables can be found as tables in the 1990 STF3 census data.

After obtaining the data from the census tables, the first step in creating the Townsend Index is to transform each area's value for each component factor *j* into a standardized Z score. The Z score for area *i* is calculated as:

$$Z_{ij} = (X_{ij} - m_j) / s_j$$

Where $X_{ij}$ is the value of component variable *j* for area *i* where $m_j$ is the mean of component *j* across all areas and $s_j$ is the standard deviation of the component variable *j* over all areas.

The second step to create this index is to sum the Z score values from each of the four components of the index.

# Tyler Street


- 86.4 % working class
- 15.6 % unemployed
- 26.5 % below poverty line
- $18,607 median household income
- 5.1 % owner-occupied homes valued >$300,000

# Mount Vernon St

(this is one home, not an apartment building)


- 26.4 % working class
- 5.4 % unemployed
- 8.0 % below poverty line
- $84,959 median household income
- 40.2 % owner-occupied homes valued >$300,000

# Analytic Methods

Our primary analytic approach for describing socioeconomic gradients by area-based socioeconomic measures has been to use geocodes to append area-based socioeconomic data to case records, to stratify these records into discrete categories based on ABSM, and to aggregate numerators and denominators over areas, within levels defined by ABSM. This method avoids the problem of unstable rates arising from small areas by assuming that cases and population denominators from areas with similar socioeconomic characteristics can be legitimately combined into the same strata. An alternative approach, which preserves the spatial information of the geocodes, is discussed in the section on multilevel analyses.

The following steps are used to generate age-standardized disease rates stratified by area-based socioeconomic measures, once the case data have been geocoded and appropriate ABSMs have been generated from census data.

• Aggregate the case data into numerators (age cells within areas/geocodes).
• Aggregate population denominator data into age cells within areas/geocodes.
• Merge the numerators and denominators with ABSMs, by area/geocode.
• Aggregate over areas into strata defined by categorical ABSM and age category.
• Generate age-standardized rates and other summary measures.

**[See the 'Tools" section for a step by step comparison of the analytic methods, the relevant task of the Case Example, and sample SAS code.]**

## Aggregating Numerator Data

Data from public health databases are typically formatted such that each record represents one person (or case report). Once these data have been geocoded, they need to be aggregated before linking to denominator and ABSM data. Before aggregating, however, one should exclude all records that are not geocoded, do not meet the case definition, or are missing data on the important covariates (e.g. age, in the case of simple age-standardized analyses; age, sex, and race/ethnicity in the case of more complex stratified analyses).

One can think of the basic unit of aggregation as a cell, defined by age and other covariates, within an area/geocode. Once aggregated, this cell within an area can be linked to a relevant population denominator. The cell contains a count of all cases within that area that meet the specified age and other covariate criteria. Since our goal is eventually to create rates, we call this count of cases the "numerator."

Example:
We intend to age-standardize in 5 broad age categories, 0-14, 15-24, 25-44, 45-64, 65+. Therefore, we need to aggregate the records in each census tract into cells defined by the corresponding ages. As an example, consider the following 23 records from census tracts 25009250500 and 25009250800.

**Before aggregating:**

| Record # | Geocode | Age of death |
|---|---|---|
| 1 | 25009250500 | <1 |
| 2 | 25009250500 | <1 |
| 3 | 25009250500 | <1 |
| 4 | 25009250500 | 17 |

| | | |
|---|---|---|
| 5 | 25009250500 | 19 |
| 6 | 25009250500 | 27 |
| 7 | 25009250500 | 38 |
| 8 | 25009250500 | 40 |
| 9 | 25009250500 | 40 |
| 10 | 25009250500 | 44 |
| 11 | 25009250800 | <1 |
| 12 | 25009250800 | <1 |
| 13 | 25009250800 | 5 |
| 14 | 25009250800 | 22 |
| 15 | 25009250800 | 24 |
| 16 | 25009250800 | 26 |
| 17 | 25009250800 | 31 |
| 18 | 25009250800 | 36 |
| 19 | 25009250800 | 36 |
| 20 | 25009250800 | 40 |
| 21 | 25009250800 | 43 |
| 22 | 25009250800 | 43 |
| 23 | 25009250800 | 43 |

**After aggregating:**

| Geocode | Age category | Number of deaths (numerator) |
|---|---|---|
| 25009250500 | 0-14 | 3 |
| 25009250500 | 15-24 | 2 |
| 25009250500 | 25-44 | 5 |
| 25009250800 | 0-14 | 3 |
| 25009250800 | 15-24 | 2 |
| 25009250800 | 25-44 | 8 |

# Aggregating Denominator Data

Denominator data at the census tract level typically come from the decennial census. In 1990, the US Census reported population counts by age in 31 categories (<1, 1-2, 3-4, 5, 6, 7-9, 10-11, 12-13, 14, 15, 16, 17, 18, 19, 20, 21, 22-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-61, 62-64, 65-69, 70-74, 75-79, 80-84, 85+). In the 1990 US Census STF3, age specific population counts were reported in table P013. Variable P0130001 gave the count of residents <1 year old, P0130002 gave the count of residents 1-2 years old, etc.

For the purposes of age standardization, these age categories need to be re-aggregated to match the age categories used for categorizing case data (numerators, above) and the age categories from the standard million reference population. Additionally, when using case data from multiple years, in order to calculate an average annual incidence rate, one needs to use a person-time denominator (population count multiplied by number of years of case data). For example, in the case of the Massachusetts all-cause mortality data, we have three years worth of cases (1989-1991). Therefore, we multiply the population count in each age category by 3.

Example:
For census tract 25009250800 in 1990, we wish to age standardize using the same five broad age categories as in the numerator example above (0-14, 15-24, 25-44, 45-64, 65+):

**Before:**

| Census variable | Ages (years) | Population count |
|---|---|---|
| P0130001 | <1 | 115 |
| P0130002 | 1-2 | 243 |
| P0130003 | 3-4 | 197 |
| P0130004 | 5 | 92 |
| P0130005 | 6 | 59 |
| P0130006 | 7-9 | 237 |
| P0130007 | 10-11 | 160 |
| P0130008 | 12-13 | 141 |
| P0130009 | 14 | 77 |
| P0130010 | 15 | 62 |
| P0130011 | 16 | 54 |
| P0130012 | 17 | 94 |
| P0130013 | 18 | 65 |
| P0130014 | 19 | 89 |
| P0130015 | 20 | 101 |
| P0130016 | 21 | 128 |
| P0130017 | 22-24 | 387 |
| P0130018 | 25-29 | 571 |
| P0130019 | 30-34 | 746 |
| P0130020 | 35-39 | 422 |
| P0130021 | 40-44 | 354 |
| P0130022 | 45-49 | 317 |
| P0130023 | 50-54 | 176 |
| P0130024 | 55-59 | 174 |
| P0130025 | 60-61 | 65 |
| P0130026 | 62-64 | 214 |
| P0130027 | 65-69 | 158 |
| P0130028 | 70-74 | 316 |
| P0130029 | 75-79 | 178 |
| P0130030 | 80-84 | 112 |
| P0130031 | 85+ | 69 |

In order to collapse these variables into the five broad age categories, we have to sum up census variables as follows:

**After:**

| Age category | Population count | | Person-time denominator (x 3 years of case data) |
|---|---|---|---|
| P0130001 | <1 | 1321 | 115 |
| P0130002 | 1-2 | 980 | 243 |
| P0130003 | 3-4 | 2093 | 197 |
| P0130004 | 5 | 946 | 92 |
| P0130005 | 6 | 833 | 59 |

# Merging numerators with denominators and ABSM

Once the numerators and denominators have the same structure (AREAKEY x AGECAT), they can be merged together, along with the ABSM data (by AREAKEY). For age cells within areas where no cases were reported, we set the numerator to zero.

Example:

**Before merging with ABSM:**

**Numerator dataset:**

| Geocode/Areakey | Age category | Number of deaths (numerator) |
|---|---|---|
| 25009250500 | 0-14 | 3 |
| 25009250500 | 15-24 | 2 |
| 25009250500 | 25-44 | 5 |
| 25009250500 | 45-64 | 7 |
| 25009250500 | 65+ | 26 |
| 25009250800 | 0-14 | 4 |
| 25009250800 | 15-24 | 3 |
| 25009250800 | 25-44 | 8 |
| 25009250800 | 45-64 | 13 |
| 25009250800 | 65+ | 132 |

**Denominator dataset:**

| Geocode/Areakey | Age category | Person-time denominator (x 3 years of case data) |
|---|---|---|
| 25009250500 | 0-14 | 4152 |
| 25009250500 | 15-24 | 1953 |
| 25009250500 | 25-44 | 3489 |
| 25009250500 | 45-64 | 1233 |
| 25009250500 | 65+ | 1212 |
| 25009250800 | 0-14 | 3963 |
| 25009250800 | 15-24 | 2940 |
| 25009250800 | 25-44 | 6279 |
| 25009250800 | 45-64 | 2838 |

| | | | |
|---|---|---|---|
| 25009250800 | 65+ | | 2499 |

**After merging with ABSM:**

| Geocode | Age category | Poverty | Numerator | Denominator |
|---|---|---|---|---|
| 25009250500 | 1 | 4 | 3 | 4152 |
| 25009250500 | 2 | 4 | 2 | 1953 |
| 25009250500 | 3 | 4 | 5 | 3489 |
| 25009250500 | 4 | 4 | 7 | 1233 |
| 25009250500 | 5 | 4 | 26 | 1212 |
| 25009250800 | 1 | 3 | 4 | 3963 |
| 25009250800 | 2 | 3 | 3 | 2940 |
| 25009250800 | 3 | 3 | 8 | 6279 |
| 25009250800 | 4 | 3 | 13 | 2838 |
| 25009250800 | 5 | 3 | 132 | 2499 |

# Aggregating OVER areas in ABSM strata

Next, in order to generate rates for categories of a specific ABSM, it is necessary to aggregate OVER areas into strata defined by AGECAT and ABSM. Numerators and denominators from census tracts with missing ABSM data for a particular ABSM are typically excluded from that analysis.

Example: In Suffolk County, Massachusetts, there are a total of 189 census tracts. We wish to examine all cause mortality rates by poverty, with poverty categorized into 4 strata (0-4.9%, 5-9.9%, 10-19.9%, and 20-100%).

| ABSM: CT Poverty | Number of census tracts |
|---|---|
| 0.0-4.9% | 10 |
| 5.0-9.9% | 37 |
| 10.0-19.9% | 56 |
| 20.0-100.0% | 83 |
| Missing poverty data | 3 |

Thus, to obtain the mortality rates in the least impoverished stratum (0.0-4.9% below poverty), we need to aggregate the cases and the population at risk OVER the ten census tracts in that stratum (preserving the age structure WITHIN each poverty stratum so that we can age standardize in the following step, below). For the next poverty stratum (5.0-9.9%) we need to aggregate the cases and the population denominator over 37 census tracts, and so on. Cases and population at risk in the three census tracts with missing poverty data are excluded from the analysis.

This yields the following table:

| ABSM: CT poverty | Age category | Numerator | Denominator |
|---|---|---|---|
| 0.0-4.9% | 0-14 | 1 | 10,608 |
| 0.0-4.9% | 15-24 | 5 | 9,984 |
| 0.0-4.9% | 25-44 | 54 | 29,190 |
| 0.0-4.9% | 45-64 | 106 | 16,710 |

| | | | |
|---|---|---|---|
| 0.0-4.9% | 65+ | 657 | 15,825 |
| 5.0-9.9% | 0-14 | 40 | 69,939 |
| 5.0-9.9% | 15-24 | 39 | 64,065 |
| 5.0-9.9% | 25-44 | 252 | 179,595 |
| 5.0-9.9% | 45-64 | 792 | 90,042 |
| 5.0-9.9% | 65+ | 4,535 | 80,916 |
| 10.0-19.9% | 0-14 | 101 | 88,989 |
| 10.0-19.9% | 15-24 | 93 | 93,147 |
| 10.0-19.9% | 25-44 | 531 | 224,793 |
| 10.0-19.9% | 45-64 | 962 | 100,479 |
| 10.0-19.9% | 65+ | 3,944 | 71,955 |
| 20.0-100.0% | 0-14 | 182 | 155,193 |
| 20.0-100.0% | 15-24 | 170 | 217,593 |
| 20.0-100.0% | 25-44 | 831 | 288,882 |
| 20.0-100.0% | 45-64 | 1,291 | 108,588 |
| 20.0-100.0% | 65+ | 3,645 | 72,720 |

# Generating Rates and Other Summary Measures/Measures of Effect

## 1. Age-standardized incidence rates

The standard practice of public health departments in reporting population rates of mortality and disease incidence is to calculate age-standardized rates, which facilitates comparisons between regions or subgroups of interest. The age-standardized rate is interpretable as the rate that would be observed in a population if that population had the same age distribution as a given reference population. Standardization by the direct method involves taking a weighted average of the age specific incidence rates observed in the area or subgroup of interest, where the weights come from a standard age distribution, such as the year 2000 standard million [1].

"Standard million" reference populations are available based on the US population age distribution for 1940, 1970, 1980, 1990, and 2000. Here we present the standard million in 11 age categories.

| **Age (years)** | **Standard million reference population** | | | | |
|---|---|---|---|---|---|
| | **Year 1940** | **Year 1970** | **Year 1980** | **Year 1990** | **Year 2000** |
| <1 | 15,343 | 17,150 | 15,598 | 12,936 | 13,818 |
| 1-4 | 64,718 | 67,265 | 56,565 | 60,863 | 55,317 |
| 5-14 | 170,355 | 200,511 | 154,238 | 141,584 | 145,565 |
| 15-24 | 181,677 | 174,405 | 187,542 | 147,860 | 138,646 |
| 25-34 | 162,066 | 122,567 | 163,683 | 173,600 | 135,573 |
| 35-44 | 139,237 | 113,616 | 113,155 | 151,095 | 162,613 |
| 45-54 | 117,811 | 114,265 | 100,641 | 101,416 | 134,834 |
| 55-64 | 80,294 | 91,481 | 95,799 | 85,030 | 87,247 |
| 65-74 | 48,426 | 61,192 | 68,775 | 72,802 | 66,037 |
| 75-84 | 17,303 | 30,112 | 34,116 | 40,429 | 44,842 |

| | | | | | |
|---|---|---|---|---|---|
| 85+ | 2,770 | 7,436 | 9,888 | 12,385 | 15,508 |

For our project, we used five broad age categories to age standardize, in order to obtain more stable rates in each age stratum, particularly for outcomes with sparse data. The relationship between our five categories and the standard eleven categories is illustrated in the table below.

| **Age in 11 categories** | **Year 2000 standard million** | **Age in 5 categories** | **Year 2000 standard million** |
|---|---|---|---|
| <1 | 13,818 | | |
| 1-4 | 55,317 | <15 | 214,700 |
| 5-14 | 145,565 | | |
| 15-24 | 138,646 | 15-24 | 138,646 |
| 25-34 | 135,573 | 25-44 | 298,186 |
| 35-44 | 162,613 | | |
| 45-54 | 134,834 | 45-64 | 222,081 |
| 55-64 | 87,247 | | |
| 65-74 | 66,037 | | |
| 75-84 | 44,842 | 65+ | 126,387 |
| 85+ | 15,508 | | |

If $cases_j$ represents the number of cases in age group *j* of the group or region of interest and $pop_j$represents the population associated with that age group, then the standardized rate $IR_{st}$for the group or region is

$$IR_{st} = \frac{\sum_j w_j \left(\frac{\text{cases}_j}{\text{pop}_j}\right)}{\sum_j w_j} = \frac{\sum_j w_j IR_j}{\sum_j w_j}$$

where $w_j$is the weight associated with category *j* in the reference (standardizing) population (e.g. the population size or the proportion of the total population). The estimated variance of the standardized rate is given by:

$$Var(IR_{st}) = \frac{\sum_j {w^2}_j \left(\frac{\text{cases}_j}{\text{pop}_j^2}\right)}{\left(\sum_j w_j\right)^2}$$

(When the $w_j s$ are proportions, then $IR_{st} = \sum_j w_j IR_j$ and $Var(IR_{st}) = \sum_j {w^2}_j \left(\frac{\text{cases}_j}{\text{pop}_j^2}\right)$).

Example:
To calculate the age-standardized all cause mortality rates in each of the four poverty strata in Suffolk County, we start with the age-specific mortality data. In each poverty stratum, the age standardized mortality rate is calculated as a weighted sum of the age-specific mortality rates, with the weights for each age stratum defined by the Year 2000 standard million.

| ABSM: CT poverty | Age category | Numerator | Denominator | Year 2000 standard million | $w_j$(weight) | $IR_j$(incidence rate per 100,000) | $IR_{st}$(age standardized rate per 100,000) |
|---|---|---|---|---|---|---|---|
| 0.0-4.9% | 0-14 | 1 | 10,608 | 214,700 | 0.215 | 9.4 | |
| 0.0-4.9% | 15-24 | 5 | 9,984 | 138,646 | 0.139 | 50.1 | |
| 0.0-4.9% | 25-44 | 54 | 29,190 | 298,186 | 0.298 | 185 | 729.7 |
| 0.0-4.9% | 45-64 | 106 | 16,710 | 222,081 | 0.222 | 634.4 | |
| 0.0-4.9% | 65+ | 657 | 15,825 | 126,387 | 0.126 | 4,151.70 | |
| 5.0-9.9% | 0-14 | 40 | 69,939 | 214,700 | 0.215 | 57.2 | |
| 5.0-9.9% | 15-24 | 39 | 64,065 | 138,646 | 0.139 | 60.9 | |
| 5.0-9.9% | 25-44 | 252 | 179,595 | 298,186 | 0.298 | 140.3 | 966.2 |
| 5.0-9.9% | 45-64 | 792 | 90,042 | 222,081 | 0.222 | 879.6 | |
| 5.0-9.9% | 65+ | 4,535 | 80,916 | 126,387 | 0.126 | 5,604.60 | |
| 10.0-19.9% | 0-14 | 101 | 88,989 | 214,700 | 0.215 | 113.5 | |
| 10.0-19.9% | 15-24 | 93 | 93,147 | 138,646 | 0.139 | 99.8 | |
| 10.0-19.9% | 25-44 | 531 | 224,793 | 298,186 | 0.298 | 236.2 | 1,014.0 |
| 10.0-19.9% | 45-64 | 962 | 100,479 | 222,081 | 0.222 | 957.4 | |
| 10.0-19.9% | 65+ | 3,944 | 71,955 | 126,387 | 0.126 | 5,481.20 | |
| 20.0-100.0% | 0-14 | 182 | 155,193 | 214,700 | 0.215 | 117.3 | |
| 20.0-100.0% | 15-24 | 170 | 217,593 | 138,646 | 0.139 | 78.1 | |
| 20.0-100.0% | 25-44 | 831 | 288,882 | 298,186 | 0.298 | 287.7 | 1,019.30 |
| 20.0-100.0% | 45-64 | 1,291 | 108,588 | 222,081 | 0.222 | 1,188.90 | |
| 20.0-100.0% | 65+ | 3,645 | 72,720 | 126,387 | 0.126 | 5,012.40 | |

## 2. Confidence intervals for directly standardized rates

Traditional confidence limits for the direct standardized rates are based on the normal distribution and require large cell counts. In our analyses, we found that they can also occasionally result in "impossible" lower limits that are less than zero. Because of this, we adopted an alternate method for calculating the confidence limits based on the inverse gamma function [2]. This method assumes that the direct standardized rate is a linear combination of independent Poisson random variables. Assuming that this linear combination also follows a Poisson distribution, the age-standardized rate $E(X) = x$ follows a gamma distribution $\Gamma(a,b)$ as follows:

$$X \backsim \Gamma\left(\frac{x^2}{v}, \frac{v}{x}\right)$$

where $x$ is the age-standardized rate ($IR_{st}$ as estimated above) and v is its variance, as described above. Converting this to the gamma distribution in its standard form, i.e. where b=1, this yields

$$\frac{X}{b} \backsim \Gamma\left(\frac{x^2}{v}, 1\right)$$

which greatly simplifies calculations. Then the lower 100(1-α) confidence limit for $\frac{x^2}{v}$ is given by

$$L\left(\frac{x^2}{v}\right) = \Gamma^{-1}\left(\frac{x^2}{v}, 1\right)(\alpha/2)$$

and the upper 100(1-α) confidence limit for $\frac{x^2}{v}$ is given by

$$U\left(\frac{x^2}{v}\right) = \Gamma^{-1}\left(\frac{(x+k_M)^2}{(v+k_M)}, 1\right)\left(1-\frac{\alpha}{2}\right)$$

where $k = k_M = \max_{j\epsilon(1,\ldots,j)}(k_j)$ is a continuity correction necessitated by using a continuous distribution to estimate confidence limits for a discrete random variable.

Increasing the number of events by 1 in an age stratum i results in a $k_j = \frac{w_j}{pop_j}$ increase in the age-standardized rate. If $k_j$ is constant for all age intervals, then $k_j = k$. However, since the $w_j$ and $pop_j$ typically vary across age strata, it is unclear what value of $k$ to use. A very conservative upper limit can be obtained by using the maximum value of $k_j = k_M$. However, following the recommendation of the NCHS, we used a close approximation that alleviates the need to calculate $k_M$:

$$U\left(\frac{x^2}{v}\right) = \Gamma^{-1}\left(\frac{x^2}{v}+1, 1\right)\left(1-\frac{\alpha}{2}\right)$$

To transform these intervals to obtain the desired confidence limits for *X*, we use $L(X) = \frac{L(x^2/v)}{x/v}$ and $U(X) = \frac{U(x^2/v)}{x/v}$.

Example:
In the following analysis of mortality due to homicide and legal intervention among hispanic women in Massachusetts, the lower confidence limits on the rate in the 5.0-9.9% poverty stratum is negative, using the traditional normal approximation method. In contrast, the lower confidence limit based on the gamma distribution yields a more reasonable confidence limit.

| ABSM: CT poverty | Rate per 100,000 | Confidence Limits Normal approximation | | "Gamma" interval | | Deaths | Person-time at risk |
|---|---|---|---|---|---|---|---|
| | | Lower | Upper | Lower | Upper | | |
| 0.0-4.9% | 0 | (0.0 | ,0.0) | (0.0 | ,9.2) | 0 | 40,182 |
| 5.0-9.9% | 3.5 | -(0.5 | ,7.5) | (0.7 | ,10.3) | 3 | 67,458 |
| 10.0-19.9% | 3.8 | (0.1 | ,7.5) | (1.0 | ,9.7) | 4 | 87,336 |
| 20.0-100.0% | 4.2 | (1.4 | ,7.0) | (1.9 | ,8.0) | 11 | 228,288 |

## 3. Confidence intervals for IRst

When the observed rate is zero (i.e. there were zero cases), the gamma method is unable to produce confidence limits for the direct standardized rates. In this situation, we adopt the following convention for the confidence limit. The lower limit is simply set to zero. For the upper limit, we assume that the number of cases (i.e. the count) follows a Poisson distribution, and use the formula for the "exact" upper confidence limit of a Poisson random variable [3]:

$$U(Y) = \frac{1}{2}\chi^{-1}_{2(y+1)df}\left(1 - \frac{\alpha}{2}\right)$$

where y is the count, i.e. zero. When α = 0.05 (i.e. for a 95% confidence limit) this simplifies to $U(Y) = \frac{\chi^{-1}_{2df}(1-\alpha/2)}{2} = 3.689$.

We can then divide this upper limit on the count by the population denominator to give the upper limit on the rate.

Example:
In the analysis of mortality due to homicide and legal intervention among Hispanic women in Massachusetts, the estimated rate in the least impoverished group is zero, since there were no deaths reported in census tracts with 0-4.9% below poverty. In the table below, the normal approximation method yields a confidence interval of (0,0) for the rate in the least impoverished group, as well (as "impossible" negative lower limits on the rates in the 5.0-9.9% poverty stratum, as we saw above). The gamma method also yields a (0,0) interval for the rate in the least impoverished group, so we have corrected the entry for the upper confidence limit as described

above. Using the "exact" upper limit on the *count* of 3.689, we divide this by the denominator (40,182) to give an upper limit of 9.2 per 100,000.

| ABSM: CT poverty | $IR_{st}$(age standardized rate per 100,000) | Confidence Limits Normal approximation | | "Gamma" interval | | Deaths | Person-time at risk |
|---|---|---|---|---|---|---|---|
| | | Lower | Upper | Lower | Upper | | |
| 0.0-4.9% | 0 | (0.0 | ,0.0) | (0.0 | ,9.2) | 0 | 40182 |
| 5.0-9.9% | 3.5 | -(0.5 | ,7.5) | (0.7 | ,10.3) | 3 | 67458 |
| 10.0-19.9% | 3.8 | (0.1 | ,7.5) | (1.0 | ,9.7) | 4 | 87336 |
| 20.0-100.0% | 4.2 | (1.4 | ,7.0) | (1.9 | ,8.0) | 11 | 228288 |

## 4. Age-standardized incidence rate difference and rate ratio

Two commonly used measures for comparing incidence rates from two different groups are the incidence rate difference (IRD) and the incidence rate ratio (IRR). The incidence rate difference compares the rates on the absolute scale, and summarizes the excess rate comparing the larger to the smaller rate. The incidence rate ratio compares the rates on a relative scale, summarizing the size of one rate relative to the other rate.

To compare two age-standardized incidence rates on the absolute scale, the age-standardized incidence rate difference ($IRD_{st}$) is the rate in one group minus the rate in the other, i.e. $IR_{st1} - IR_{st0}$. The variance of this age-standardized incidence rate difference is simply the sum of the estimated variance of the two age-standardized rates [4],

$$Var\,(IRD) = Var\,(IR_{st1}) + Var\,(IR_{st0})$$

To compare age-standardized rates from two different groups or regions on the relative scale, the age-standardized incidence rate ratio ($\mathrm{IRR}_{st}$) is simply $\mathrm{IR}_{st1}/IR_{st0}$. Confidence intervals can be calculated using the variance estimator [4]:

$$Var\,[log\,(IRR_{st})] = \frac{Var\,(IR_{st1})}{{IR_{st1}}^2} + \frac{Var\,(IR_{st0})}{{IR_{st0}}^2}$$

Example:
To compare the age-standardized incidence rates in the most and least impoverished census tracts in Suffolk County, we start with the age-specific data for these two strata (note: for ease of presentation, we present variances in scientific notation in the table below):

| ABSM: CT Poverty | Age category | Numerator | Denominator | $w_j$ (weight) | $\mathrm{IR}_j$(age specific rate) | Var($\mathrm{IR}_j$) (variance of the age specific rate) | $\mathrm{IR}_{st}$(age standardized rate) | Var($\mathrm{IR}_{st}$) (variance of the age standardized rate) |
|---|---|---|---|---|---|---|---|---|
| 0.0-4.9% | 0-14 | 1 | 10,608 | 0.2147 | 0.000094 | 8.89E-09 | | |
| 0.0-4.9% | 15-24 | 5 | 9,984 | 0.1386 | 0.000501 | 5.02E-08 | 0.007297 | 6.76E-08 |
| 0.0-4.9% | 25-44 | 54 | 29,190 | 0.2982 | 0.00185 | 6.34E-08 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.0-4.9% | 45-64 | 106 | 16,710 | 0.2221 | 0.006344 | 3.80E-07 | | |
| 0.0-4.9% | 65+ | 657 | 15,825 | 0.1264 | 0.041517 | 2.62E-06 | | |
| 20.0-100.0% | 0-14 | 182 | 155,193 | 0.2147 | 0.001173 | 7.56E-09 | | |
| 20.0-100.0% | 15-24 | 170 | 217,593 | 0.1386 | 0.000781 | 3.59E-09 | | |
| 20.0-100.0% | 25-44 | 831 | 288,882 | 0.2982 | 0.002877 | 9.96E-09 | 0.010193 | 1.77E-08 |
| 20.0-100.0% | 45-64 | 1,291 | 108,588 | 0.2221 | 0.011889 | 1.10E-07 | | |
| 20.0-100.0% | 65+ | 3,645 | 72,720 | 0.1264 | 0.050124 | 6.89E-07 | | |

The **age-standardized rate difference** is simply 1,019.3 per 100,000 – 729.7 per 100,000 = 289.6 per 100,000 (or, in scientific notation, 2.896 x 10-3).

Using the formula above, we calculate the variance of $\mathrm{IRD}_{st}$.

$$Var\left(IRD_{st}\right) = 6.76 \times 10^{-8} + 1.77 \times 10^{-8} = 8.54 \times 10^{-8}$$

Then the lower and upper confidence limits are derived as follows:

$$L_{IRDst} = 2.896 \times 10^{-3} - \left(1.96 * \sqrt{8.54 \times 10^{-8}}\right) = 0.002323$$

$$U_{IRDst} = 2.896 \times 10^{-3} - \left(1.96 * \sqrt{8.54 \times 10^{-8}}\right) = 0.003469$$

or, expressed per 100,000, 232.2 to 346.9 per 100,000.

The **age-standardized rate ratio** is simply 1,019.3 per 100,000/729.7 per 100,000 = 1.40.

Using the formula above, we calculate the variance of $\log(IRR_{st})$:

$$Var\left[\log\left(IRR_{st}\right)\right] = \frac{6.76 \times 10^{-8}}{0.007297^2} + \frac{1.77 \times 10^{-8}}{0.010193^2} = 0.001441$$

Then the lower and upper confidence limits are derived as follows:

$$L_{IRR_{st}} = \exp\left[\log\left(1.40\right) - 1.96\sqrt{0.001441}\right] = 1.30$$

$$U_{IRR_{st}} = \exp\left[\log\left(1.40\right) - 1.96\sqrt{0.001441}\right] = 1.50$$

## 5. Relative Index of Inequality (RII)

Comparisons of socioeconomic gradients based on categorical ABSM may be complicated by differences in the population distributions of area-based socioeconomic measures. For example, it may be expected that the

classifications producing smaller groups at the margins would lead to larger incidence rate ratios, comparing the most deprived to the most affluent, because finer discrimination of extremes of socioeconomic position is achieved. The relative index of inequality (RII) has been proposed as a measure which explicitly addresses this problem [5 – 7]. Assuming ordinality of the ABSM categories, the RII is calculated by regressing the incidence rate in each ABSM category on the total proportion of the population that is more deprived in the socioeconomic hierarchy. Because the RII combines information about the magnitude of the socioeconomic gradient with information about the distribution of the socioeconomic variable in the population, it can be conceptualized as a measure of "total population input".

In practice, this latter quantity is represented by the cumulative distribution function (cdf). We approximate the cdf for the jth level of a given ABSM by summing the proportion of the population represented by the categories $\mathrm{ABSM}_1$,…, $\mathrm{ABSM}_{j-1}$, and adding one-half the proportion of the population represented by the category $\mathrm{ABSM}_j$.

Example:
In order to calculate the RII for poverty and all cause mortality in Massachusetts, we begin by calculating the approximate cumulative distribution function as follows:

| ABSM: CT poverty | Population denominator | Proportion | Formula | Approximate cdf |
|---|---|---|---|---|
| 0.0-4.9% | 7,626,117 | 0.423 | 0.2115 | 0.211 |
| 5.0-9.9% | 5,508,912 | 0.305 | 0.5755 | 0.576 |
| 10.0-19.9% | 2,782,194 | 0.154 | 0.805 | 0.805 |
| 20.0-100.0% | 2,120,208 | 0.118 | 0.941 | 0.941 |

In order to compare RII meaningfully across groups with differing age composition, we developed an age-standardized RII, standardized to the year 2000 standard million, as follows. Let $observed_{ij}$ be the observed number of cases in the *ith* age group and the *jth* category of ABSM, and $pop_{ij}$ be the population at risk in the corresponding category. First, we calculate the age-standardized rate $IR_{st}$ in each stratum j defined by ABSM, as described above. For each stratum *j*, we estimate the expected number of cases in stratum j, $\mathrm{expected}_j$, by multiplying the age-standardized rate $IR_{st}$ by the population denominator, $pop_j = \Sigma_i pop_{ij}$. We determine the "marginal" cumulative distribution function, *cdf(ABSM$_j$)*, of the ABSM over the entire population, as noted above.

Example:
The column of red numbers shows the expected number of cases in each poverty stratum.

| ABSM: CT Poverty | IRst (age standardized rate per 100,000) | Observed deaths | Population denominator | Expected deaths | Approximate cdf |
|---|---|---|---|---|---|
| 0-4.9% | 757 | 57,256 | 7,626,117 | 57,731.70 | 0.211 |
| 5-9.9% | 840.3 | 52,583 | 5,508,912 | 46,291.70 | 0.576 |
| 10-19.9% | 915.9 | 27,730 | 2,782,194 | 25,482.00 | 0.805 |
| 20-100% | 1,035.30 | 17,842 | 2,120,208 | 21,950.70 | 0.941 |

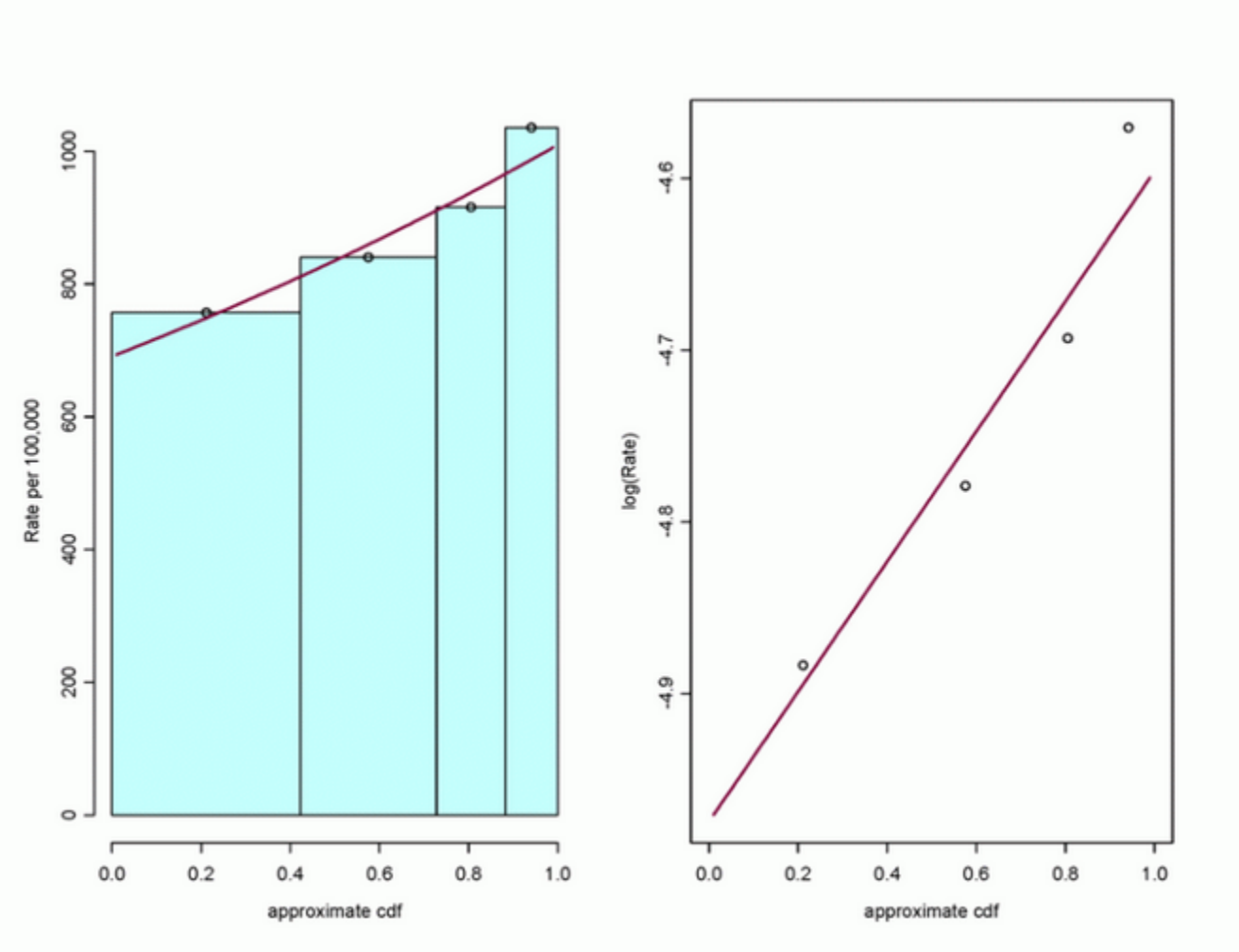
To calculate the age-standardized RIIst, we fit the following Poisson model for the expected cases:

$$expected_{ij} \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = log(pop_{ij}) + \beta_0 + \beta_1 * cdf(ABSM_j)$$

Exponentiation of the $\beta_1$ yields the RII, which is interpretable as an incidence rate ratio comparing the rates in the bottom to the top of the socioeconomic hierarchy. A larger RII indicates a greater the degree of inequality across a socioeconomic hierarchy, which may be due to a steep socioeconomic gradient or large inequalities in the distribution of the ABSM itself.

Example:
Fitting this model to the data presented above yields a $\beta_1$ of 0.379. Exponentiating this, we obtain an RII of 1.46. In the figures below, we can see how the RII for poverty is obtained. In the left figure, the height of light blue bars represents the all cause mortality rate per 100,000 in each of the four poverty strata (0-4.9, 5-9.9, 10-19.9, 20-100%), with width of bars proportional to population size of poverty stratum (in order from least to most impoverished). Open circles are plotted along the x-axis at the interpolated midpoints of each bar, approximating the cumulative distribution function of CT level poverty. The solid line represents fitted RII line. In the left figure, this line is not a straight line since the fitted line comes from a Poisson model. The right figure shows the plotted points and fitted RII line on the log scale, where the line is truly straight.




## 6. Population Attributable Fraction

The population attributable fraction (PAF) is a useful summary measure for characterizing the public health impact of an exposure on population patterns of health and disease. It is defined as "the fraction of all cases (exposed and unexposed) that would not have occurred if exposure had not occurred." [8]For a polytymous exposure, the population attributable fraction is a weighted sum of the attributable fractions for each level of the exposure, with the weights defined by the case fractions (number of exposed cases divided by overall number of cases):

$$PAF = CF_1 \times \frac{RR_1 - 1}{RR_1} + CF_2 \times \frac{RR_2 - 1}{RR_2} + \ldots + CF_j \times \frac{RR_j - 1}{RR_j}$$

In order to aggregate multiple PAFs over several age strata i=1,…,I, note that

$$PAF_{agg} = \frac{\sum_i \text{excess number of cases}}{\text{number of cases}}$$

$$= \frac{\sum_i \text{number of cases} \times \frac{\text{excess number of cases}}{\text{number of cases}}}{\sum_i \text{number of cases}}$$

$$= \frac{\sum_i \text{number of cases} \times PAF_1}{\sum_i \text{number of cases}}$$

that is, a weighted average of stratum specific PAFs, with the number of cases in each age stratum as weights.

Example:
To calculate the population attributable fraction of all cause mortality due to poverty, we begin by tabulating the cases and population person-time at risk in each poverty stratum j within each age group i. Within each age group, the case fraction $\mathrm{CF}_{ij}$is the number of cases in that poverty stratum, divided by the total number of cases within the age group. The incidence rate ratio $\mathrm{IRR}_j$for a particular poverty stratum, relative to the reference category of the least impoverished group, is calculated by dividing the rate in that poverty stratum by the rate in the least impoverished group. For each age stratum, we calculate a separate age-specific PAF, as seen in the column of red numbers in the table below. These age-specific PAFs range from 5% to 23%.

| Age category (i) | ABSM: CT poverty (j) | Cases | Person-time denominator | Rate per 100,000 | Case Fraction ( $\mathrm{CF}_{ij}$) | Incidence rate ratio ( $\mathrm{IRR}_{ij}$) | Population attributable fraction ( $\mathrm{PAF}_i$) |
|---|---|---|---|---|---|---|---|
| 0-14 | 0-4.9% (reference) | 303 | 727,947 | 41.6 | 40.7% | 1.00 | 0.1626 |
| | 5.0-9.9% | 253 | 461,958 | 54.8 | 34.0% | 1.32 | |
| | 10.0-19.9% | 113 | 206,214 | 54.8 | 15.2% | 1.32 | |
| | 20.0-100.0% | 75 | 100,716 | 74.5 | 10.1% | 1.79 | |
| | Total cases: | 744 | | | | | |
| 15-24 | 0-4.9% (reference) | 377 | 510,645 | 73.8 | 40.6% | 1.00 | 0.0506 |
| | 5.0-9.9% | 323 | 349,518 | 92.4 | 34.8% | 1.25 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 10.0-19.9% | 152 | 179,928 | 84.5 | 16.4% | 1.14 | |
| | 20.0-100.0% | 76 | 153,273 | 49.6 | 8.2% | 0.67 | |
| | Total cases: | 928 | | | | | |
| | 0-4.9% (reference) | 1,569 | 1,201,002 | 130.6 | 34.7% | 1.00 | |
| | 5.0-9.9% | 1,392 | 873,072 | 159.4 | 30.7% | 1.22 | |
| 25-44 | 10.0-19.9% | 933 | 405,366 | 230.2 | 20.6% | 1.76 | 0.2266 |
| | 20.0-100.0% | 633 | 200,457 | 315.8 | 14.0% | 2.42 | |
| | Total cases: | 4,527 | | | | | |
| | 0-4.9% (reference) | 5,314 | 763,464 | 696.0 | 39.7% | 1.00 | |
| | 5.0-9.9% | 4,429 | 461,451 | 959.8 | 33.1% | 1.38 | |
| 45-64 | 10.0-19.9% | 2,287 | 191,934 | 1,191.6 | 17.1% | 1.71 | 0.2210 |
| | 20.0-100.0% | 1,369 | 82,674 | 1,655.9 | 10.2% | 2.38 | |
| | Total cases: | 13,399 | | | | | |
| | 0-4.9% (reference) | 19,470 | 376,002 | 5,178.2 | 38.8% | 1.00 | |
| | 5.0-9.9% | 17,784 | 314,181 | 5,660.4 | 35.4% | 1.09 | |
| 65+ | 10.0-19.9% | 8,734 | 146,091 | 5,978.5 | 17.4% | 1.15 | 0.0725 |
| | 20.0-100.0% | 4,248 | 63,594 | 6,679.9 | 8.5% | 1.29 | |
| | Total cases: | 50,236 | | | | | |

To aggregate these PAFs across age strata, we weight the contribution of each age stratum by the proportion of cases in that age stratum. As seen in the table below, this results in an aggregated population attributable fraction $\mathrm{PAF}_{agg}$ of 11%.

| Age category (i) | Cases | Population attributable fraction ( $\mathrm{PAF}_i$) | $\rightarrow$ | Aggregated population attributable fraction ( $\mathrm{PAF}_{agg}$) |
|---|---|---|---|---|
| 0-14 | 744 | 0.1626 | (744*0.1626 + 928*0.0506 + 4527*0.2266 + 13399*0.2210 + 50236*0.0725)/ 69834 | =0.1116 |
| 15-24 | 928 | 0.0506 | | |
| 25-44 | 4,527 | 0.2266 | | |
| 45-64 | 13,399 | 0.221 | | |
| 65+ | 50,236 | 0.0725 | | |

Total cases: 69,834

## REFERENCES

1. Breslow NE, Day NE (eds). Statistical Methods in Cancer Research, Vol. II: The Design and Analysis of Cohort Studies. Oxford, UK: Oxford University Press, 1987.

2. Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard; National Vital Statistics Reports: Vol 37, No. 3. Hyattsville, MD: National Center for Health Statistics, 1998.

3. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Statistics in Medicine 1997;16:791-801.

4. Rothman KJ, Greenland S. Modern Epidemiology. 2nd Edition. Philadelphia: Lippincott-Raven, 1998.

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

6. Wagstaff A, Paci P, van Doorslaer E. On the measurement of inequalities in health. Soc Sci Med 1991;33:545-57.

7. 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;52:153-60.

8. JA Hanley, A heuristic approach to the formulas for population attributable fraction. J Epidemiol Community Health 2001;55:508-514.

# Multi-level Modeling

It is well known that there are substantial area variations in mortality rates in the U.S. However, the presence of area differences in mortality does not necessarily mean that area matters. Area variations in mortality can be observed due to a number of reasons some of which may be due to characteristics that relate to areas and others that relate to the characteristics of the individuals who live in these areas. Disentangling the two sources of variation (e.g.: individual and area) in mortality is therefore vital to distinguishing area differences from the difference that area makes. Such an approach to examining area variations in mortality, consequently, entails describing the patterning and causes in mortality variations, which in turn, requires answering the following empirical questions preferably in a sequential manner.

Before we outline the questions, it is worth asking what role could places or areas play in influencing mortality (and indeed other health outcomes). Pure locational attributes of an area (e.g., altitude, proximity to coast) or environmental aspects of an area (e.g., levels of air pollution) or structural attributes of an area (e.g., residential segregation, labor markets, population density) or collective social aspects of an area (e.g., proportion of poor in an area, proportion population that has less than high school education) are some concrete elements along which area variations in mortality may get patterned. Indeed, the different examples mentioned above need not be mutually exclusive. Thus, an examination of area variations and area-based explanations to these variations could be addressed by answering the following questions:

- First, how does the total variation in mortality get partitioned across the individual and area levels?

- Second, how much of the variation in mortality that is attributable to areas is influenced by the characteristics of individual residents who live in these areas?
- Third, does the magnitude of variation in mortality that is attributable to areas differ for different population groups? For instance, is the area-attributable variation in mortality greater for blacks as compared to whites?
- Fourth, to what extent do area-based characteristics account for the area-attributable variation in mortality, in whites and blacks, for example?
- Fifth, what is the systematic relationship between area-based characteristics and mortality, and does this relationship systematically differ across different population sub-groups?

Answering these types of questions requires adopting a multilevel statistical modeling approach (also known as hierarchical, mixed and random-effects, covariance components or random-coefficient regression). These techniques have provided researchers one possible framework for incorporating and understanding the role of areas and context while studying mortality variations. The key advantage of this approach is, therefore, in analyzing, "why some areas are more likely to experience higher levels of mortality, while taking into account of why some individuals (independent of which area they live) are more likely to die".

The use of multilevel statistical techniques is especially pertinent under the following circumstances:

The first is when the individual health outcome measure (or group-specific prevalence) are anticipated to be clustered with the source of clustering being a geographic area, such as block-groups or/and census-tracts and the interest is in ascertaining the relative importance of the different levels for the outcome. This is particularly relevant for public health departments as they provide a clue about the level at which actions occur. The assessment of what level matters the most for the outcomes can be done unconditionally (not adjusting for covariates) and conditionally (adjusted for covariates).

The second situation that necessitates the use of multilevel methods is when the exposure is measured at multiple levels and the interest is in evaluating the relative importance of a same ABSM at different levels (e.g.: establishing whether the block-group poverty has a larger effect than the census-tract poverty).

Finally, multilevel methods offer a bridge between statistical modeling and descriptive map-based presentations. Since the specific census-tracts and block-groups identifiers are intrinsic to the analytical design, it is possible to develop conditional statistical maps showing how different places are doing on a particular health outcome and importantly whether the "geography of health" differs for different population sub-groups. This provides a useful means to monitor health inequalities that is conditional on a range of important socioeconomic characteristics. Technical benefits also flow from utilizing this perspective. There of course are serious substantive issues (such as "naming" and "shaming" places) as well technical issues (such as instability in intrinsically small areas with less population; mismatch of outcome measure with the denominator information) that need to be considered given the immediate appeal of maps. While strategies drawing upon "empirically bayes" modeling (utilized widely within the multilevel models) or smoothing may bring certain technical solutions, issues of mapping for small areas in particular are complex and substantive.

While this approach is gaining usage in public health research, given the relative complexity of these modeling strategies it is yet to become a part of the mainstream public health surveillance and monitoring. At the same time, the reasons to empirically evaluate the above questions are compelling. For instance, patterns of all cause mortality are likely to be shaped by a complex constellation of compositional and contextual factors that may conceivably vary for different population subgroups, as suggested, for example, by different leading causes of death for different racial/ethnic groups. An investigation of the racial/ethnic heterogeneity in geographic variation in mortality can give insight into the relative importance of compositional and contextual effects to mortality experienced by different racial/ethnic populations. For example, if the geographic variation in

mortality rates for a specific group is large, this suggests that geographically varying contextual factors may be of particular importance in shaping mortality risk for this population. Conversely, if the geographic variation in mortality rates is low for a particular group, it suggests that contextual factors are of relatively less importance in shaping overall mortality risk for that population.

The subject of modeling area-related effects – through measuring the area-attributable variation and through identifying area-based characteristics – is intrinsically multilevel and this note outlined the sort of questions and motivations that could underlie investigations of variation in health and mortality.

Multilevel models may now be implemented using a variety of software packages including SAS, STATA, R and MLwiN. The Center for Multilevel Modeling website provides a list of these software packages at http://multilevel.ioc.ac.uk/softrev/index.html

***For fundamental texts, see:***


- Goldstein H. Multilevel statistical models. 2nd ed. London: Arnold, 1995.
- Longford N. Random coefficient models. Oxford: Clarendon Press, 1993.
- Raudenbush S, Bryk A. Hierarchical linear models: applications and data analysis methods. Thousand Oaks: Sage, 2002.


***For applied introductions to multilevel statistical models, see:***


- Hox J. Multilevel analysis: techniques and applications. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
- Leyland AH, Goldstein H. Multilevel modelling of health statistics. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd., 2001.
- Snijders T, Bosker R. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications, 1999.
- Subramanian SV, Jones K, Duncan C, 2003, Multilevel methods for public health research, in Kawachi I, Berkman L. Eds. Neighborhoods and Health, New York: Oxford University Press, 65-111.


***For hands-on tutorial, see:***


- Browne WJ. MCMC estimation in MLwiN. London: Centre for Multilevel Modelling, Institute of Education, 2002.
- Rasbash J, Browne W, Goldstein H, Yang M, Plewis I, Healy M, Woodhouse G, Draper D, Langford I, Lewis T. A user's guide to MLwiN, Version 2.1. London: Multilevel Models Project, Institute of Education, University of London, 2000.

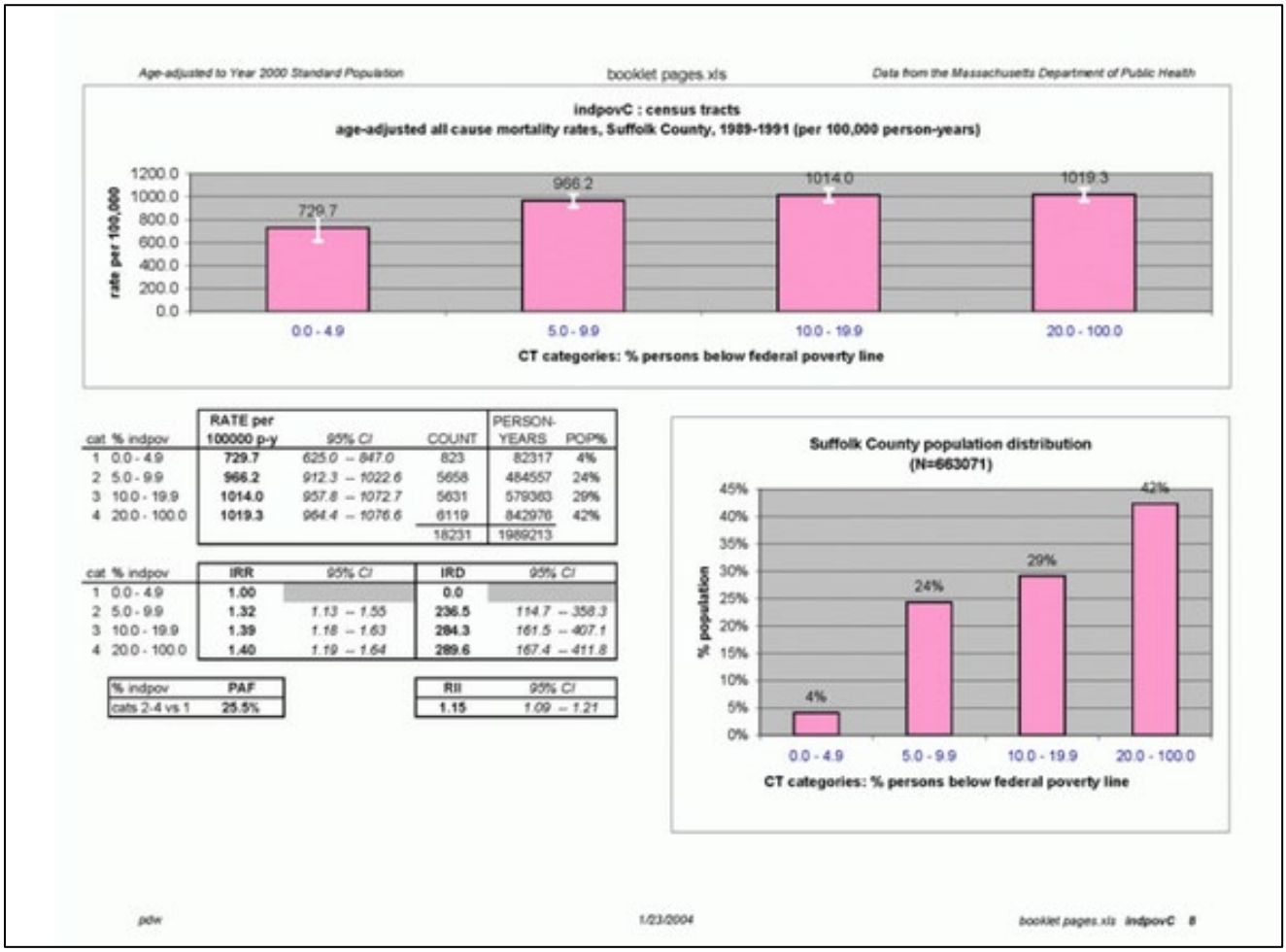

***For issues related to mapping see:***


- Elliott P, Wakefield J, Best N, Briggs D (eds). Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press, 2000.
- Maantay J. Mapping environmental injustices: pitfalls and potential of geographic information systems in assessing environmental health and equity. Environ Health Perspect 2002; 110 (suppl 2):161-171.
- Monmonier M. How to Lie with Maps. 2nd ed. Chicago: University of Chicago Press, 1996.
- Monmonier M. Cartographies of Danger: Mapping Hazards in America. Chicago: University of Chicago Press, 1997.
- Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol Rev 1999 21:143-161.

- 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.

# Visual Display

A visual display of the data can offer a dramatic and succinct representation of the socioeconomic disparities in your data. Often, graphical representations provide a means of communicating key features of the data, and can enhance summary presentations of data in tabular form. For example, as part of our project, we produced a series of booklets – one for each health outcome, at each level of geography. Each page of the booklet summarized the standardized incidence rates, rate ratio, rate differences, relative index of inequality, and population attributable fraction, for each ABSM, and was supplemented by a visual display of the incidence rates in each category of the ABSM, and the population distribution of the ABSM. Figure 1 below shows a sample booklet page summarizing the analysis of all cause mortality in Suffolk County, Massachusetts, by CT poverty. (This is the same analysis presented in our case example). Similar pages could be constructed to summarize all cause mortality by % working class, % less than high school education, etc.




| cat | % indpov | RATE per 100000 p-y | 95% CI | COUNT | PERSON-YEARS | POP% |
|---|---|---|---|---|---|---|
| 1 | 0.0 - 4.9 | 729.7 | 625.0 -- 847.0 | 823 | 82317 | 4% |
| 2 | 5.0 - 9.9 | 966.2 | 912.3 -- 1022.6 | 5658 | 484557 | 24% |
| 3 | 10.0 - 19.9 | 1014.0 | 957.8 -- 1072.7 | 5631 | 579363 | 29% |
| 4 | 20.0 - 100.0 | 1019.3 | 964.4 -- 1076.6 | 6119 | 842976 | 42% |
| | | | | 18231 | 1989213 | |

| cat | % indpov | IRR | 95% CI | IRD | 95% CI |
|---|---|---|---|---|---|
| 1 | 0.0 - 4.9 | 1.00 | | 0.0 | |
| 2 | 5.0 - 9.9 | 1.32 | 1.13 -- 1.55 | 236.5 | 114.7 -- 358.3 |
| 3 | 10.0 - 19.9 | 1.39 | 1.18 -- 1.63 | 284.3 | 161.5 -- 407.1 |
| 4 | 20.0 - 100.0 | 1.40 | 1.19 -- 1.64 | 289.6 | 167.4 -- 411.8 |

| % indpov | PAF |
|---|---|
| cats 2-4 vs 1 | 25.5% |

| RII | 95% CI |
|---|---|
| 1.15 | 1.09 -- 1.21 |

Figure 1. Booklet Page

Maps of Census derived ABSMs can also give a dramatic visual representation of how socioeconomic conditions are distributed geographically. In the figure below, we mapped CT level poverty in Suffolk County, MA, using ArcView/ArcGIS.

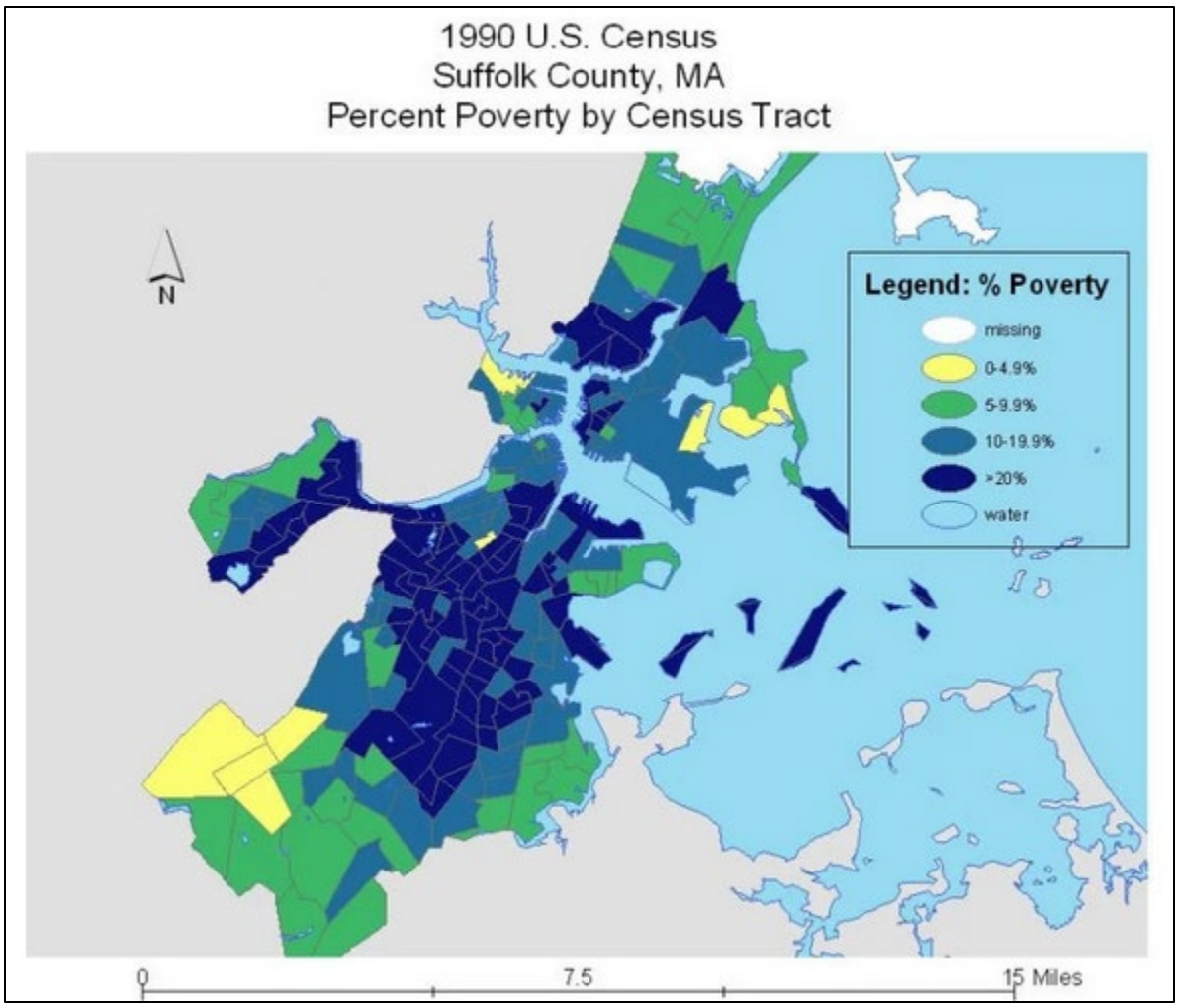



Figure 2. Map of Suffolk County, MA Poverty

Mapping of disease rates at the census tract level can present complications, however, because rates for small areas are often unstable due to small numbers. For this reason, and because our focus was on area-based socioeconomic disparities in health across all of Massachusetts and Rhode Island, rather than within specific census tracts, we explicitly chose not to map disease rates as part of this project.

Another way of displaying the data is shown in Figure 2 from the Introduction. In these graphs, which we newly apply to routinely collected U.S. state health department data1-3, the width of each bar is proportional to the size of the population in the specified socioeconomic statum4. We created these graphs in S-plus. Here's an example of this mode of graphing applied to our case example.

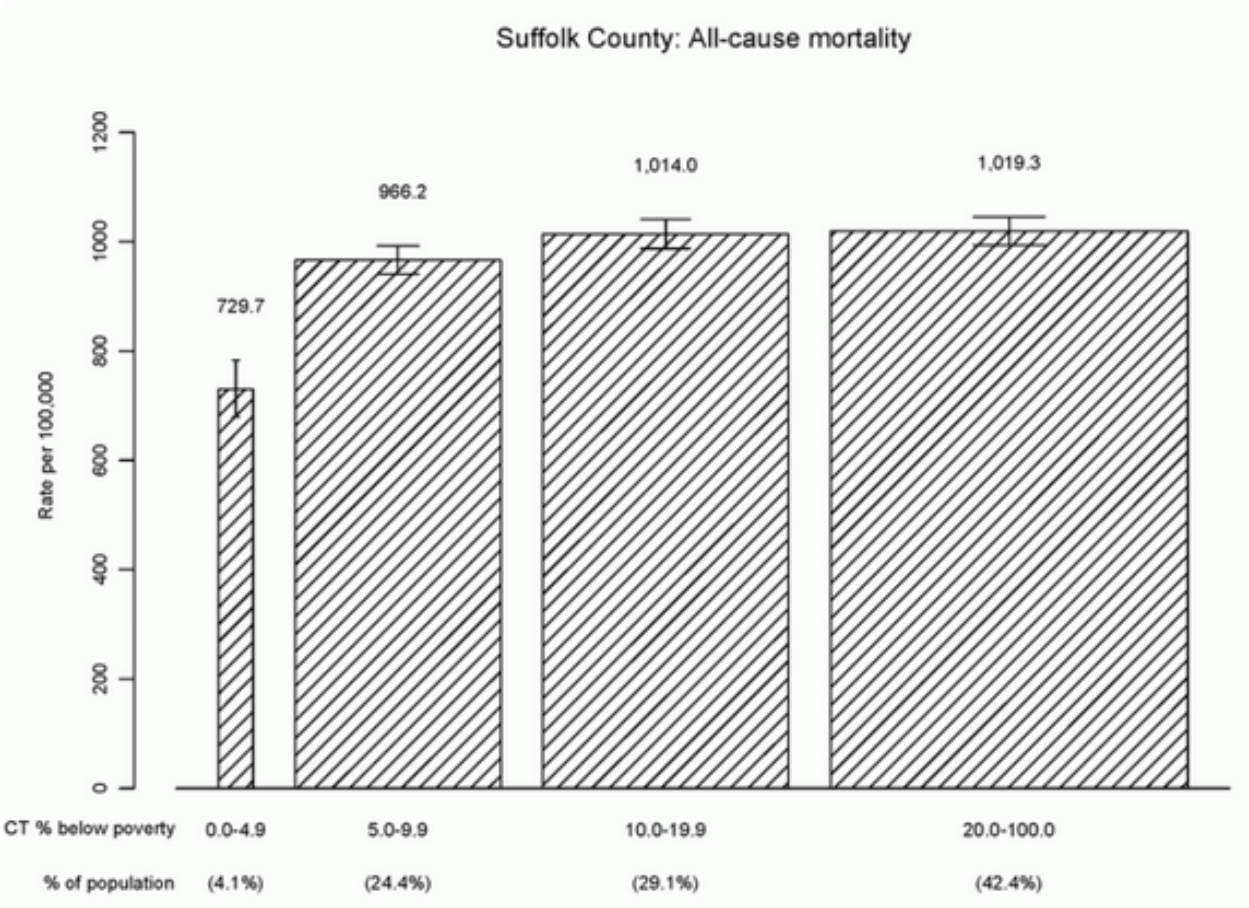



Figure 3. Graph of Suffolk County, MA Poverty by All-cause mortality

**REFERENCES**

1.Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. 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 2002; 156:471-482. https://pubmed.ncbi.nlm.nih.gov/12196317/

2. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). J Epidemiol Community Health 2003; 57:186-199. https://pubmed.ncbi.nlm.nih.gov/12594195/

3. Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian S. Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures–The Public Health Disparities Geocoding Project (US). Public Health Rep 2003; 118:240-260. https://pubmed.ncbi.nlm.nih.gov/12766219/

4. Wagstaff A, Paci P, van Doorslaer E. On the measurement of inequalities in health. Soc Sci Med 1991;33:545-57.

# Tools

The **Case Example** is an opportunity for programmers and data managers to try out the techniques we describe on a test dataset, drawn from all-cause mortality cases in Suffolk County, MA, from 1989 to 1991. We provide test datasets, a step-by-step description of the programming tasks, sample SAS code, and examples of the resulting output.

To facilitate further research on socioeconomic gradients in health with respect to our recommended area-based socioeconomic measure (CT poverty), we have made available **Census Tract Level Poverty Data** for ALL census tracts in the United States, for 1980, 1990, and 2000.

# Case Example

We created a **Case Example** as an opportunity for you to experiment with our methodology. This example draws on all cause mortality data from Suffolk County, Massachusetts, between 1989 and 1991. You'll have a chance to analyze these data by census tract poverty to see the socioeconomic gradient in mortality in this county. We've divided the case example into clearly defined tasks to highlight the process of moving from raw data to summary measures of the socioeconomic disparity.

In order to complete the exercise, you will also need these raw data files and SAS program.

# U.S. Census Tract Poverty Data

To see more about US census tract poverty data and the cut-point for poverty areas defined by the US census as >=20% below poverty, see the US Census Bureau Changes in Poverty Rates and Poverty Areas Over Time: 2005-2019.

To facilitate monitoring of area-based socioeconomic disparities in health data, we have extracted census tract level poverty data for ALL census tracts in the United States, for 1980, 1990, and 2000, from the U.S. Census. We have made these available as comma-delimited text files for each year of the decennial Census (1980, 1990, and 2000) with two fields per record.

The first field is the 11-digit areakey (i.e. the geocode) which uniquely identifies the census tract:

Digits 1-2 = State code
Digits 3-5 = County code
Digits 6-11 = Census Tract code (often used with a decimal point:xxxx.xx)
Digit 12 = Blockgroup code

The second field is the percent of people in the census tract living below the federally defined poverty line:
To calculate the proportion of persons below poverty for this region, we sum all categories P1170001 to P1170024 to get the denominator, and sum categories P1170013 to P1170024 to get the numerator, and then simply divide this numerator by the denominator:

$$(P1170001 + \ldots + P1170012) / (P1170001 + \ldots + P1170024)$$

Once you have geocoded your own data, you can merge your data with these CT-level poverty data, by areakey.

Poverty 1980
Poverty 1990
Poverty 2000

**Note:** If you open these data files in Excel, the first column (11-digit areakey) may not be properly displayed. Simply widen the column to see the full areakey.

# Glossary

**ABSM** see "Area-based socioeconomic measure"

**address cleaning** The process of taking an original address and retaining only key elements of that address (building number, street and street type), as well as correcting spelling errors and standardizing abbreviations.

**age stratum** One category of age in a series of age categories.

**American Community Survey** A new national survey administered by the US Census Bureau that provides yearly data on states and counties between the decennial censuses and which, by 2008, should provide these data for census tracts as well. For more information see http://www.census.gov/acs/www/ .

**area** A geographic region whose boundaries may be defined socially, topographically, or ecologically (singly or in combination).

**area-based measure** see "area-based socioeconomic measure"

**area-based socioeconomic measure** A specifically defined measure that is used to characterize the socioeconomic conditions of an area (as opposed to the socioeconomic position of individuals); for example, percent of persons living below poverty.

**block group** "A subdivision of a census tract, generally containing between 600 and 3,000 people, with an optimum size of 1,500 people. Most block groups were delineated by local participants as part of the U.S. Census Bureau's Participant Statistical Areas Program. It is the lowest level of the geographic hierarchy for which the U.S. Census Bureau tabulates and presents sample data. (from Appendix A. Census 2000 Geographic Terms and Concepts. http://www.census.gov/geo/www/tiger/glossry2.pdf)

**Carstairs Index UK** Composite deprivation measure, created by summing standardized Z scores from area-based data on percent crowding, percent male unemployment, percent no car ownership, and percent low social class.

**case record** see case report

**case report** Data on an individual that indicates the incidence or prevalence of a morbidity or mortality outcome.

**cdf** see cumulative distribution function

**cell** A basic unit of aggregation based on the cross-classification of a number of categorical variables. For example, all cases occurring among women ages 40-44 in a given census tract are aggregated into a single cell defined by gender, age, and area.

**census geography** A scheme of classification of areas used by the U.S. census. For example, census tract and block group are both types of areas by which data are classified in U.S. census data.

**census tract** "A small relatively permanent statistical subdivision delineated by local participants as part of the U.S. Census Bureau's Participant Statistical Areas program. When first delineated they are designed to be relatively homogenous with respect to population characteristics, economic status and living conditions. They

average in size between 1,500 and 8,000 people, with an optimum size of 4,000 people. The geographic size varies considerably depending on population density. (from Appendix A. Census 2000 Geographic Terms and Concepts. http://www.census.gov/geo/www/tiger/ glossry2.pdf)

**census variable** Items of data organized by the U.S. Census bureau. Data for these variables is structured in the form of census tables, that may include one or more census variables.

**class** see social class

**comma-delimited file** A text file format where data fields are separated by commas. The Microsoft Excel file extension for this type of data is .csv .

**composite index** see composite measure

**composite measure** A measure that combines information on more than one component variable. For example, the Townsend index consists of percent unemployment, percent renters, percent not owning a car, and percent crowding.

**compositional factors** Attributes of areas that derive from the characteristics of individuals.

**construct** A theoretical concept or idea.

**contextual factors** Attributes of areas that derive from structural or social characteristics of the area.

**CT** see census tract

**cumulative distribution function** For a given value, the area under the probability function up to that value (i.e. cdf(x) = Pr[X<=x]). When calculated as part of deriving the relative index of inequality, the cumulative distribution function of an area-based socioeconomic measure (ordered from most affluent to most deprived) for a given value can be interpreted as the proportion of the population who are more affluent.

**denominator** There are two definitions of denominator that depend on the measure being calculated. For calculating rates, the denominator is the amount of person-time observed during which time cases were eligible to occur. For calculating ABSMs, the denominator is the total number of persons in an area for which the ABSM was measured.

**deprivation** "Deprivation can be conceptualized and measured, at both the individual and area level, in relation to: material deprivation, referring to 'dietary, clothing, housing, home facilities, environment, location and work (paid and unpaid), and social deprivation, referring to rights in relation to 'employment, family activities, integration into the community, formal participation in social institutions, recreation and education' "(from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**direct age standardization** A method for adjusting a population rate for age, yielding the hypothetical rate that would have been observed if the population being studied had the same age distribution as an externally defined standard population. In direct standardization, stratum specific rates are multiplied by weights derived from a standard reference population, and summed to yield a summary rate. Rates standardized to the same external standard may be meaningfully compared to examine differences that are not due to age.

**ecosocial theory** A theory that seeks to "integrate social and biological reasoning and a dynamic, historical and ecological perspective to develop new insights into determinants of population distributions of disease and

social inequalities in health." The core concepts for ecosocial theory include 1. embodiment, 2. pathways to embodiment, 3. cumulative interplay between exposure, susceptibility, and resistance, and 4. accountability and agency. (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**etiologic period** The duration of time over which a disease develops, referring to the time from an initial exposure to the time at which the outcome caused by this exposure occurs.

**exact confidence limits** Exact confidence limits that do not rely on a normal approximation. We used exact confidence limits to calculate confidence intervals when the rate was zero.

**gamma confidence intervals** Confidence intervals for the direct standardized rate based on the gamma distribution. A practical consequence of using gamma confidence intervals is that confidence intervals for rates will not cross zero. For more details see Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Statistics in Medicine 1997;16:791-801

**gender** "A social construct regarding culture-bound conventions, roles and behaviors for, as well as relationships between and among, women and men and boys and girls." (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**geocoding** The assignment of a numeric code to a geographical location

**geographical information systems** Technology based systems that combine layers of geographic data to offer a greater understanding of the characteristics of places.

**georesult** see MatchCode

**Gini** A measurement of inequality that ranges between 0 and 1, which is the ratio of the area under the Lorenz curve to the area under the diagonal on a graph of the Lorenz curve. A value of one would indicate complete inequality of distribution, while a 0 indicates no inequality.

**GIS** see geographical information systems

**incidence rate** The number of events divided by the person-time at risk.

**incidence rate difference** The absolute difference between two incidence rates. The incidence rate among the exposed proportion of the population, minus by the incidence rate in the unexposed portion of the population, gives an absolute measure of the effect of a given exposure.

**incidence rate ratio** The ratio of two incidence rates. The incidence rate among the exposed proportion of the population, divided by the incidence rate in the unexposed portion of the population, gives a relative measure of the effect of a given exposure.

**index of local economic resources** A composite index based on "white collar employment, unemployment rate, and median family income, developed for use at the county level" (see Casper ML, Barnett E, Halverson JA, Elmer GA, Braham VE, Majeed ZA, Bloom AS, Stanley S. Women And Heart Disease: An Atlas Of Racial And Ethnic Disparities In Mortality. Office for Social Environment and Health Research, West Virginia University, Morgantown, WV, 1999.) Data for the three component variables are ranked into deciles, and then summed.

**indirect age standardization** A method for adjusting a population rate for age, yielding the hypothetical rate that would have been observed if the population being studied had the same age distribution as an externally defined standard population. Indirect standardization is based on deriving an expected number of events using an externally defined standard population, and contrasting this value to the observed number of events in the population being studied. The expected number of events is derived by multiplying the stratum-specific counts in the study population by stratum-specific rates from a standard population. The ratio of total observed events to the number expected is the standardized mortality (or morbidity) ratio (SMR). The indirect standardized rate is calculated by multiplying the SMR by the crude rate from the standard population.

**injury due to legal intervention** Includes injuries inflicted by the police or other law-enforcing agents, including military on duty, in the course of arresting or attempting to arrest lawbreakers, suppressing disturbances, maintaining order, and other legal action.

**lifecourse perspective** "Refers to how health status at any given age, for a given birth cohort, reflects not only contemporary conditions but embodiment of prior living circumstances, in utero onwards" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**MatchCode** An indicator of which address elements determined the geocode, thus giving an indication of the accuracy of the geocode (also called "georesult" by some companies).

**material deprivation** see deprivation

**multilevel analysis** Analyses that conceptualize and analyze associations at multiple levels, e.g., employ individual- and area-based data in relation to a specified outcome. These analyses typically entail the use of variance components models to partition the variance at multiple levels, and to examine the contribution of factors measured at these different levels to the overall variation in the outcome.

**non-fatal weapons related injuries** A category of injury that includes intentional and unintentional non-fatal gun and stabbing injuries.

**numerator** There are two definitions of numerator that depend on the measure calculated. For calculating rates, the numerator is the number of events observed. For calculating ABSMs, the numerator is the number of persons or households in an area with the socioeconomic characteristic of interest.

**occupational class** A measurement of socioeconomic position based upon job characteristics. One example is the British Registrar General's Social Class scheme, based on skill. This was replaced in 2001 by an occupational measure based on job relations, the National Statistics Socio-Economic Classification system (NS-SEC); related, in this study "working class" occupations were conceptually defined as those as those employing non-supervisory employees (and for the ABSM "working class" measure, were operationally defined as those census occupational categories comprised chiefly of working class occupations).

**operational definition** A description of a variable in terms of how the variable is actually measured.

**person-time** The sum of the time at risk for all persons in a population.

**Poisson** model A regression model used for count data.

**population attributable fraction** The theoretical reduction of incidence that would be expected if the entire population had the same level of exposure as a specified referent group (which could be a group with low or no exposure).

**poverty** "To be impoverished is to lack or be denied adequate resources to participate meaningfully in society" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**poverty area** In the US, the federal criteria for being a "poverty area" is to be an area with a 20% or more of the population below the poverty line.

**poverty** line A poverty threshold that takes into account household size and age composition and intended to indicate an income level below which subsistence needs are not met. The poverty line in the US is based on a value of three times the cost of the economy food basket in 1963, adjusted for inflation. See: "How the Census Bureau Measures Poverty (Official Measure)" at: http://www.census.gov/hhes/poverty/ povdef.html

**public health surveillance system** A structure that facilitates the continuous and systematic collection of descriptive information for monitoring the health of populations (from Buehler, Chapter 22: Surveillance, in Rothman and Greenland, Modern Epidemiology, 2nd edition, 1998, p 435-457).

**race/ethnicity** "A social, not biological, category, referring to social groups, often sharing cultural heritage and ancestry, that are forged by oppressive systems of race relations, justified by ideology, in which one group benefits from dominating other groups, and defines itself and others through this domination and the possession of selective and arbitrary physical characteristics (for example, skin color)" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**rate difference** see incidence rate difference

**rate ratio** see incidence rate ratio

**relative index of inequality** A summary measure of "total population impact" that takes into account both the socioeconomic gradient in the outcome, as well as the population distribution of the socioeconomic variable. The RII is interpretable as the ratio of the rate in the theoretically most deprived segment of the population, compared to the rate in the theoretically least deprived segment.

**RII** see relative index of inequality

**SEP** see socioeconomic position

**sex** "A biological construct premised upon biological characteristics enabling sexual reproduction" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**social class** "Refers to social groups arising from interdependent economic relationships among people" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**social deprivation** see deprivation

**socioeconomic position** "An aggregate concept that includes both resource-based and prestige-based measures, as linked to both childhood and adult social class position" (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**socioeconomic status** A term referring to prestige-based measures of socioeconomic position, as determined by rankings in a social hierarchy (from Krieger N. A Glossary for Social Epidemiology, J Epidemiol Community Health 2001; 55:693-700.)

**spatiotemporal** Of, relating to, or existing in both space and time.

**spatiotemporal mismatch** A mismatch of data derived from different sources that arises because of (1) inconsistency of boundaries between data sources and/or (2) inconsistency of timeframe between data sources.

**S-Plus** Commercially available software for data modeling and statistical analysis. A similar version of this software named R is available for free under a GNU General Public License at www.r-project.org.

**STF3 table** A table of census data from the Summary Tape File 3 of the US census (until 2000, when replaced by Summary File 3) that provides full and sample count data for socioeconomic and other census variables down to the census tract and block group level.

**Townsend Index UK** Deprivation measure consisting of a standardized Z score combining data on percent crowding, percent unemployment, percent no car ownership, and percent renters.

**transpose** To reverse the orientation of a matrix, so that the values across the rows become the values down the columns, and the values of the columns become the values across the rows.

**wealth** Conceptually, wealth refers to accumulated assets. An ABSM to capture wealth is operationalized from census data as percent of owner-occupied homes worth more than 400% of the median value of owned homes.

**year 2000 standard million** The distribution of the U.S. population into 11 age categories, based on the US population structure in the Year 2000. (see: Anderson RN, Rosenberg HM. Age standardization of death rates: implementation of the year 2000 standard. National Vital Statistics Reports, Vol 37, no. 3. Hyattsville, MD: National Center for Health Statistics, 1998.)

**ZCTA** see "Zip code tabulation area"

**ZIPcode** "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" (from Appendix A. Census 2000 Geographic Terms and Concepts).

**ZIPcode tabulation area** A statistical geographic area that approximates the delivery area for a U.S. Postal service Zip code. This approximation replaces the Zip code areas used by the Census Bureau in conjunction with the 1990 and earlier censuses.(from Appendix A. Census 2000 Geographic Terms and Concepts.)

**Z-score** Also referred to as Z-ratio or Z-value, it is equal to a value of X minus the mean of X, divided by the standard deviation.

# Who We Are*

Nancy Krieger, PhD, our Principal Investigator, is a Professor of Social Epidemiology in the Department of Social and Behavioral Sciences, American Cancer Society Clinical Research Professor, and Chair of the Interdisciplinary Concentration on Women, Gender, and Health, at the Harvard T.H. Chan School of Public Health. Dr. Krieger's work focuses on social inequalities in health. She is a social epidemiologist, with a background in biochemistry, philosophy of science, history of public health, and involvement as an activist in issues involving social justice, science, and health. Her work involves: (a) etiologic studies of social inequalities in health, (b) methods for improving monitoring of social inequalities in health, and (c) development of theoretical frameworks to guide work on understanding and addressing social determinants of health.

Pamela D. Waterman, MPH, is our Project Director, based in the Department of Social and Behavioral Sciences at the Harvard T.H. Chan School of Public Health. In addition to helping keep the project on track, she served as our primary liaison with the staff of the Departments of Health and also handled all matters related to the geocoding process, including the evaluation of geocoding accuracy. She also played a major role in producing our data booklets and designed and produced the web-based and CD-ROM versions of our monograph.

Jarvis Chen, ScD, is a Research Scientist in the Department of Social and Behavioral Sciences at the Harvard T.H. Chan School of Public Health. He is a social epidemiologist and serves as the Senior Statistical Programmer for this study. In addition to developing and implementing virtually all of the programming, he also generated the graphical displays of the data.

David Rehkopf, PhD, at the time of his involvement in this project, was a Research Fellow at the Department of Society, Human Development, and Health at the Harvard School of Public Health. He presently (2017) is an Assistant Professor of Medicine at Stanford University. His research interests included investigations of income effects on health as well as the development of publicly available tools to assist in studies of population health. As a Graduate Research Assistant on this project, he assisted at all levels of data analyses – from running SAS programs to formatting output.

S V Subramanian, PhD, is Professor in the Department of Society, Human Development, and Health at the Harvard School of Public Health. He is a medical geographer with extensive experience in multi-level modeling in analysis of public health data and provided expert advice for our project's multi-level modeling of mortality and morbidity rates.

* descriptions of study teams were updated in the original on-line version of the 2004 monograph in 2017 and are retained as such here; the weblinks for each team member are to current websites (as of June 2024); Pam Waterman retired in January 2024.


Onwards!!
¡Hacia adelante!
Avante!!
En avant!!
In avanti!!
εμΠρος
衛