**Consultant's Guidance Document for CSTE utilization of the Public Health Disparities Geocoding Project Health Disparities Monitoring Methodology**

BACKGROUND: In 1998, to address the lack of socioeconomic data in most United States public health surveillance systems, and the associated inability to monitor socioeconomic disparities in health, Dr. Nancy Krieger, Professor, Harvard School of Public Health, created the Public Health Disparities Geocoding Project (PHDGP) (funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR)). The project's aim was to determine what area-based socioeconomic measure at which level of geography would be most apt for monitoring socioeconomic inequalities in health. After carefully weighing geocoding options and testing for geocoding accuracy[1], the PHDGP used a geocoding service to geocode public health surveillance data and linked these data to US Census-derived area-based socioeconomic measures (ABSMs), thereby enabling computation of rates stratified by ABSM and thus, a method for monitoring socioeconomic inequities in health.

Thirteen different datasets from the Massachusetts Department of Public Health and the Rhode Island Department of Health were analyzed. These datasets comprised 7 different health outcomes spanning the lifecourse, including: mortality (all cause and cause-specific), low birthweight, childhood lead poisoning, sexually transmitted infections, tuberculosis, cancer incidence, and non-fatal weapons-related injuries; together, they totaled approximately 1, 000,000 records.

Eighteen different ABSMs were constructed, including single-item measures and various indices, which covered 6 domains of socioeconomic position: occupational class, income, poverty, wealth, education, and crowding. ABSMs were required to meet two basic requirements: that they (a) meaningfully summarized important aspects of the area's socioeconomic conditions, and (b) employed socioeconomic data that could be compared over time and across regions.

The key finding of the research was the ABSM that was most apt for monitoring socioeconomic disparities in health was poverty (% of persons living below the federally defined poverty line); and the level of geography that was best for monitoring disparities was the census tract. The census tract poverty ABSM consistently detected the expected socioeconomic gradients in health across a wide range of health outcomes among both the total population and diverse racial/ethnic groups. Of the 18 studied ABSMs, poverty was also the ABSM most easily interpretable to health department staff. Geocoding to the census tract level as compared to blockgroup level resulted in more complete geocoding, i.e., street addresses that lacked a house number were often geocodable when it was confirmed that the entire street fell within a single census tract, noting that because of

the larger size of census tracts, more streets were entirely contained within census tracts compared to blockgroups. Additionally, because it was, and still is, a unit that could be aggregated up to administrative, and in some cases, neighborhood levels, the census tract is a unit that yields considerable weight in decisions regarding social, economic, and public health policy.

Using ZIPcodes as a unit of geographic analyses was strongly discouraged due to substantial variability of geographic area and population coverage. For example, one ZIPcode can refer to an entire county in a very rural area as well as to a single large office building in a very dense urban area. Because ZIPcodes are administered by the US Postal Service, rather than the US Census Bureau, they are not always compatible with census geography and are often changed between censuses. This can lead to difficulties in determining the appropriate population denominators for these shifting areas.

This project demonstrated that area-level socioeconomic position can be used effectively to measure and monitor socioeconomic disparities in health, as such and also in conjunction with data on racial/ethnic health

disparities. Over the past 10 years since the methodology was developed, individual health departments have implemented this methodology, including the Connecticut Emerging Infections (EIP) Program[2], the Washington State Department of Health[3], the Virginia Department of Health[4,5], and researchers at various academic institutions including the Institut national de santé publique du Québec who performed similar analyses using mortality data and Canadian Census data[6]. Informed by the PHDGP, US cancer registries have geocoded their records to the census tract level and used the poverty ABSM[7,8]. To date, however, the PHDGP methodology has not yet been used to set a national standard for monitoring US socioeconomic disparities in health.

In 2010, a working group of New York Department of Health and Mental Hygiene epidemiologists was commissioned to discuss possible standard measures for monitoring socioeconomic disparities in health in New York City, with a focus on adapting the work of the PHDGP. Although they noted a few challenges to employing the PHDGP methodology, the working group recommended that the background work done by the PHDGP research team be accepted, and that the PHDGP methodology be routinely applied to DOHMH surveillance data. In May 2011, members of the Council of State and Territorial Epidemiologists (CSTE) reviewed the PHDGP methodology and recommended its use by all member organizations. In conjunction with members of the PHDGP team, this consultants' guidance document has been prepared by members of the PHDGP research team

to facilitate use of the methodology by all CSTE member organizations to monitor socioeconomic disparities in health.

STEP ONE: Deciding which health outcome to analyze.

The PHDGP process can be applied to any health outcome. Results of analyses of the 7 health outcomes that were originally used for the project have all been published[9-12]. A nationally notifiable disease may be preferred for CSTE implementation of this methodology because all states collect the required data, and thus there is the potential for comparison among states. Health outcomes with a well-known racial/ethnic disparity might also be a good choice to showcase the methodology, as might be a health outcome of particular interest to the local constituency, or other health outcomes covered by population-based registries (e.g., cancer registries).

It is recommended, however, that jurisdictions exercise caution when using datasets with a high percentage of missing data, i.e., data missing on any variable required for the analysis with missing values for more than 20% of the observations. In some cases, missing data techniques such as multiple imputation can be used to "fill in" the missing data and obtain valid estimates under certain assumptions. We recommend that this kind of analysis be done under the guidance of a statistician. However, we acknowledge that it is possible that not all agencies will have the resources to perform the sophisticated analyses imputation would require.

For many jurisdictions, mortality data may serve as a feasible starting point for this project. Mortality data have the strength of generally being particularly robust datasets, with large numbers of observations and fairly complete data. All-cause mortality might be a good choice for initial analyses because it cuts across all diseases, noting that it is also possible to restrict to cause-specific analyses and also age-specific mortality (e.g., premature mortality, or infant mortality).

STEP TWO: Geocoding the health outcome addresses (numerators)

Many agencies currently geocode health outcome data regularly. However agencies likely utilize different methods. One agency may send addresses out to a private company to be geocoded while another agency may retain a full-time geocoder on staff. It is impractical to expect all agencies to geocode addresses using exactly the same protocol for many reasons, including staffing, access to software, budgetary constraints and confidentiality issues. Thus, rather than standardizing the method of geocoding, we recommend standardizing geocoding accuracy.

**Consultant's Guidance Document for CSTE utilization of**
**the Public Health Disparities Geocoding Project**
**Health Disparities Monitoring Methodology**

There are a number of ways that addresses from health outcome data can be geocoded. By 'geocoding' we mean determining the exact geographic location of the address associated with each health outcome record, and having done so, (a) using that information to determine which U.S. Census tract the address falls within, and (b) appending the 11-digit code for that census tract to the health outcome record. Geocoding can be done by sending addresses out to a geocoding company, using a software program, or submitting addresses to a geocoding website. Whichever method is used, it is vital that all regulations regarding confidentiality be strictly followed when geocoding addresses which fall under HIPAA guidelines. Regardless of the method chosen, it is strongly recommended that a separate data file that contains only the address fields and a unique identifier be constructed that can be used to link the geocodes back to the original health outcome file once geocoding has been completed.

To ensure accuracy of geocoding – however it is performed – it is recommend that a subset of addresses, e.g., 100, be tested against results obtained using a different geocoding service or against results obtained using a program with verified accuracy. One free, universally accessible, and reputable source is the American Factfinder Census Tract Locator [http://factfinder2.census.gov/]. Note that the census tract designation provided by the American Factfinder site will not be presented in the more familiar, concise 11-digit format; rather it will be presented in list form which will include the census tract minus place-holding zeroes.

The four geocoding services that the PHDGP tested for accuracy in 1999 scored between 44 and 84% accuracy[1]. Subsequent to this test, a subset of addresses geocoded by Company A for the actual PHDGP were tested again and scored 94% accuracy. Testing of geocoding performed by PHDGP staff using ArcGIS in 2010 also resulted in a 94% accuracy rate (unpublished data). Based on these results, the PHDGP recommends an accuracy rate of at least 94%.

Two types of geocodes that should <u>not</u> be used are: (1) any geocode that is assigned based on ZIPcode centroid (Post Office Boxes fall in this category); and (2) geocodes that are assigned with a low level of certainty, e.g., a score of less than the minimum match score of 72 as set by ArcGIS. Other geocoding software programs will likely have a comparable preset cutpoint for acceptable matches. Geocodes assigned by a commercial geocoding service that are assigned based on anything other rooftop or street address are most likely assigned based on either ZIPcode or ZIP4 centroid. Although it is possible that some ZIPcodes and ZIP4 areas are completely contained within a census tract, unless verified on a case-by-case basis, these geocodes should not be used.

STEP THREE: Extracting Census population counts (denominators)

The U.S. Census Bureau recommends that population counts from the decennial censuses, as opposed to the American Community Survey (ACS) 5-year population estimates, be used for analyses such as that proposed by CSTE for its demonstration project, and both the CSTE and the PHDGP concur with that recommendation.

The ACS surveys a sample of each state (approximately 1 per every 8 households) and projects the total population count, thus the estimate is only as good as the projection. In contrast, the decennial census surveys the entire population. Five-Year population estimates from the American Community Survey (ACS) may eventually be used, but at the time of this recommendation, they were not available for testing. Thus, we recommend that decennial census population counts be used for these analyses. Further, because the ACS poverty data (see below) are aggregated using the 2000 Census boundaries, 2000 decennial population counts must be used for the denominators for the current CSTE project and for all analyses utilizing the PHDGP methodology until ACS data aggregated to the 2010 Census boundaries are released.

For analyses with health data preceding the 2010 census, population data from the appropriate decennial censuses will need to be used. Population counts from the 1990 decennial census, for example, should be used as denominators when computing rates for health outcome data from 1985-1994, from the 2000 census for health outcome data from 1995-2004, and from the 2010 census for health outcome data from 2005- 2014. In the 1990 U.S. Census, the variable for age-specific total population counts is reported in table P013. (Variable P0130001 gives the count of residents <1 year old, P0130002 gives the count of residents 1-2 years old, etc.) In the 2000 U.S. Census, the total population counts are in table P001001, and in the 2010 census the total population variable is QT-P1.

STEP FOUR: Constructing area-based socioeconomic measures, i.e., poverty

As demonstrated by the PHDGP, the poverty area-based socioeconomic measure is a useful metric for monitoring socioeconomic disparities in health because it is: (a) robust, e.g., appropriate to use across multiple health outcomes, over diverse time periods, and (b) readily understandable. Although additional area-based socioeconomic and sociodemographic variables may eventually warrant consideration and analysis by CSTE member organizations, the census tract poverty ABSM constitutes a suitable common starting point. This measure is expressed as "% of persons living below the poverty level" (which is far easier to understand than

other commonly utilized measures e.g., "% living at 200% below poverty"). Based on its empirical findings, the PHDGP recommends specific cutpoints for % poverty: 0–4.9%, 5.0–9.9%, 10.0–19.9%, and >20%.

Analyses using this methodology by the New York City Department of Health and Mental Hygiene suggests that the cutpoints used by the PHDGP might not work for all groups, as some regions may require more categories to illuminate granular data, either because of high poverty rates or high concentrations of affluence. However, the need for comparability among regions must be kept in mind, noting in particular that the cutpoint of greater than 20% living in poverty aligns with the programmatic standard definition of a federally defined poverty area and also federally defined medically underserved area.

We therefore recommend that first time users of this methodology use the poverty variable and the PHDGP cutpoints as a starting point, with 2 more levels added within the <20% poverty stratum (20.0 – 29.9%, 30.0 – 39.9%, and >40%) to allow for more detail if needed. More advanced users can further adjust the cutpoints *within* stratum, while still allowing for aggregation back to the four recommended levels for comparisons across jurisdictions.

Percent poverty for each census tract in the United States has been calculated by the PHDGP research team using data from the 1990 and 2000 decennial censuses and the 2005-2009 ACS. These data are freely available for download from the CSTE website.

STEP FIVE: Merging datasets & analyses

Assuming that the health data are formatted such that each record represents one person (or case report), the geocoded health outcome data will need to be aggregated before they can be linked to the denominator data and to the poverty variable. If the health outcome that is being analyzed is one that is typically age-standardized, then both the numerator and denominator must be aggregated into the same age categories for age-standardization, e.g. 0-14, 15-24, 25-44, 45-64, 65+. This can be done using the same programming normally used by health department staff to aggregate for age-standardization, with the important caveat that aggregation must take place *within* individual census tracts. (If the health outcome is not one that is typically age-standardized, then the data need not be aggregated by age.)

Age-specific numerator and denominator data must be calculated for each census tract to allow for calculating age-standardized rates for each census tract. This will also allow for aggregation of data across census tracts in the same socioeconomic stratum, so as to permit calculation of age-specific and age-standardized rates for each stratum.

Thus, for example, to know the mortality rate of persons ages 25-44 in impoverished census tracts, data are required on the N of deaths among persons ages 25-44 in each census tract labeled as impoverished (e.g., poverty rate >=20%) and also the N of persons ages 25 to 44 in each such census tract. These age-specific mortality rates can be reported for each impoverished census tract, and they also can be aggregated across all impoverished census tracts, e.g, the on-average mortality rate of persons ages 25-44 living in impoverished census tracts is XXX per 100,000.

Using the below example of 5 age groupings, once the numerator data have been aggregated, a data file should now exist that has 5 observations per census tract, each representing the number of persons/cases in that age stratum that reside in that census tract.

Before aggregating:

| Record # | Census Tract | Age at death |
|---|---|---|
| 1 | 25009250500 | <1 |
| 2 | 25009250500 | <1 |
| 3 | 25009250500 | <1 |
| 4 | 25009250500 | 17 |
| 5 | 25009250500 | 19 |
| 6 | 25009250500 | 27 |
| 7 | 25009250500 | 38 |
| 8 | 25009250500 | 46 |
| 9 | 25009250500 | 40 |
| 10 | 25009250500 | 66 |
| 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 | 46 |
| 21 | 25009250800 | 47 |
| 22 | 25009250800 | 53 |
| 23 | 25009250800 | 68 |

After aggregating:

| Census Tract | Age category | Number of deaths (numerator) |
|---|---|---|
| 25009250500 | 0-14 | 3 |
| 25009250500 | 15-24 | 2 |
| 25009250500 | 25-44 | 3 |
| 25009250500 | 45-64 | 1 |
| 25009250500 | 65+ | 1 |
| 25009250800 | 0-14 | 3 |
| 25009250800 | 15-24 | 2 |
| 25009250800 | 25-44 | 4 |
| 25009250800 | 45-64 | 3 |
| 25009250800 | 65+ | 1 |

Similarly, the denominator data must be aggregated such that a data file exists with 5 observations per census tract, each representing the total number of persons in that age stratum residing in that census tract. Once the numerator and denominator have the same age structure, they can be merged together by areakey AND by age stratum.

In order to generate rates for poverty as categorized by the four poverty strata recommended by the PHDGP, all census tracts in the jurisdiction that fall into the individual stratum must be combined. All census tracts that have a poverty level of 0–4.9%, must be combined, all census tracts that have a poverty level of 5.0-9.9% must be combined, etc.

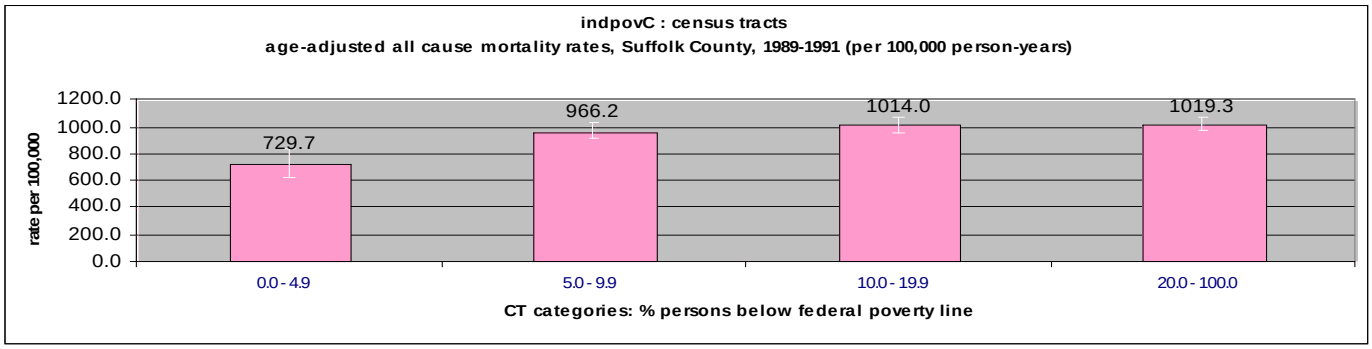
The poverty data downloaded from the CSTE website contains, in addition to the raw percentage of persons living in poverty, a variable that denotes poverty stratum according to the PHDGP guidelines, i.e., 1=0-4.9%, 2=5-9.9%, 3= 10.0-19.9%, 4=>20%. This file can now be merged with the merged numerator and denominator file, thus appending a variable that will allow aggregation of health outcome cases by poverty level as well as calculation of health outcome stratified by poverty level.

STEP SIX: Showing output

Simple histograms can be used to display output from these analyses, however it is strongly recommended that text guiding the interpretation of the data be routinely included on all output.

Example Output:

Recommended title: "Socioeconomic inequalities in [Health Outcome] rates: Comparing the burden of [Death/Disease/Disability] among persons who live in Census Tracts with fewer versus more socioeconomic resources (measured by the US poverty level)."



Recommended accompanying text: "People's social context strongly influences their risk of disease and death. The data in this chart document the association between economic resources and health status. Groups subjected to social and economic deprivation typically have worse health than people who are more economically and socially privileged – and the gaps in rates between these groups point to ill-health and death that could be prevented by reducing economic inequality."

NEXT STEPS: Guidance for conducting, evaluating, and disseminating implementation of the PHDGP methodology

First, it is recommended that any single state, local, or territory health department or health agency (or any corresponding multi-state, local, or territory consortia), who seek to implement the PHDGP should send a 1-page summary of their proposed protocol, including geocoding methodology and ABSMs to be used, to the CSTE and PHDGP, for review and advice. The PHDGP has agreed to serve, *pro bono*, in this advisory capacity, so as to help ensure rigorous and accurate implementation of the protocol. The summary should be sent with the understanding that a 2-4 week turnaround time (depending on time of year) is required for sending feedback.

Second, once the methodology has been implemented, it is important that all reports be transparent regarding the proportion of cases geocoded with precision to the census tract level (including how this may vary by other reported demographic characteristics, e.g., age, gender, and race/ethnicity), and also discuss how findings compare to relevant prior publications. At issue is both the magnitude of the estimate and its uncertainty (i.e., 95% confidence interval). For example, if considerable selection bias has occurred, e.g., due to a high proportion of records being ungeocodable with adequate precision to the census tract level (and with this proportion not randomly distributed in the population, but instead concentrated in certain groups, e.g., those whose economic constraints results in greater residential instability and hence use of PO boxes), this will affect the magnitude of observed socioeconomic disparities. Consequently, in addition to reporting on the percent of cases geocoded with precision to the census tract level, each report should systematically compare its results regarding the observed magnitude of socioeconomic disparities for the selected health outcomes to: (a) prior publications of the PHDGP, and (b) temporally relevant studies documenting the magnitude of socioeconomic disparities for the selected outcome for the relevant geographic area, whether using individual or census tract level socioeconomic data. Explicit discussion of potential biases affecting results, including consideration of whether they would lead to under- or over-estimation of the magnitude of socioeconomic disparities in health, and also their contribution to racial/ethnic health disparities, should be a standard section of any report, as should discussion of the study

strengths, e.g., the importance of evaluating the magnitude of socioeconomic disparities in health within and also across racial/ethnic groups.

Finally, the PHDGP recommends that, prior to publication (or, in the case of peer-reviewed articles, prior to submission), it be sent for methodologic review the final draft of any publication (e.g., report or document, whether hard-copy or web-based, or peer-reviewed scientific article) seeking to set national standards using the PHDGP methodology for use by state, territorial, and local health departments and health agencies. As per the offer for initial review of the proposed protocol, this review will be conducted *pro bono*, and a lead time of 2-4 weeks for requesting feedback will likewise be necessary.

REFERENCES:

1. 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.
2. Yousey-Hindes KM, Hadler JL. Neighborhood Socioeconomic Status and Influenza Hospitalizations Among Children: New Haven County, Connecticut, 2003-2010. Am J Public Health, 2011;101:1785-1789.
3. "The Health of Washington State Supplement: a statewide assessment addressing health disparities by race, ethnic group, poverty and education." September 2004. http://www.doh.wa.gov/HWS (accessed on September 16, 2011).
4. The Virginia Department of Health Epidemiology Profile 2007. http://www.vdh.virginia.gov/epidemiology/DiseasePrevention/Profile2007.htm (accessed on September 16, 2011).
5. The 2008 Virginia Health Equity Report. http://www.vdh.state.va.us/healthpolicy/2008report.htm (accessed on September 16, 2011).
6. Pampalon R, Hamel D, Gamache P. A comparison of individual and area-based socio-economic data for monitoring social inequalities in health. Health Reports 2009:20(4) available at: http://www.statcan.gc.ca/pub/82-003-x/2009004/article/11035/key-cle-eng.htm (accessed on September 16, 2011).
7. 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 Publication No. 03-0000.; available at: http://seer.cancer.gov/publications/ses/ses_monograph.pdf ; (accessed: Sept 27, 2011).
8. Harper S, Lynch J. *Methods for Measuring Cancer Disparities: Using Data Relevant to Healthy People 2010 Cancer-Related Objectives.* NCI Cancer Surveillance Monograph Series, Number 6. Bethesda, MD: National Cancer Institute, 2005. NIH Publication No. 05-5777; available at: http://seer.cancer.gov/publications/disparities/; (accessed: Sept 27, 2011).
9. 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.
10. 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.

11. 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.
12. 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.