

Anticipated Meeting Outcome: draft method guidance document that can be used by some state/local public health agencies to apply to select nationally notifiable condition(s) to analyze for health disparities.

OVERVIEW OF THE HARVARD PUBLIC HEALTH DISPARITIES GEOCODING PROJECT (PHDGP) (1R01HD36865-01; PI: NANCY KRIEGER) <http://www.hsph.harvard.edu/thegeocodingproject/>

PAMELA D. WATERMAN, MPH _ Project Director, PHDGP, Harvard School of Public Health

- In 1998, to address the lack of socioeconomic data in most public health surveillance systems, and the associated inability to monitor socioeconomic disparities in health, the Public Health Disparities Geocoding Project (PHDGP) Research team, led by Dr. Nancy Krieger, launched a project 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 accuracy, the PHDGP geocoded public health surveillance data, and linked them 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.
- 13 datasets were analyzed from the Massachusetts Department of Public Health and the Rhode Island Department of Health that comprised 7 different health outcomes (mortality [all cause and cause-specific], low birthweight, childhood lead poisoning, sexually transmitted infections, tuberculosis, and non-fatal weapons-related injuries) and totaled approximately 1, 000,000 records.
- 18 different ABSMs were constructed 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.
- Key findings: **Census tract poverty level** (% of persons living at or below the federal poverty level in a census tract) was the ABSM most apt for monitoring socioeconomic disparities in health. It consistently detected the expected socioeconomic gradients in health across a wide range of health outcomes both in the total population and within diverse racial/ethnic groups. Geocoding to the census tract level vs. blockgroup level yielded maximal returns in percentage matched, and, because census tract is a unit that can be aggregated up to legislative and larger neighborhood levels, it is a unit that yields considerable weight in decisions regarding political and public health policy. Poverty was the ABSM most easily interpretable to health department staff. Using ZIP codes as a unit of geographic analyses was strongly discouraged due to substantial variability of geographic area coverage and shifts in ZIP code designated areas between times of census population counts.
- Output: Many options for displaying socioeconomic gradients in health were explored, with the primary formats being choropleth maps and histograms designed by Dr. Jarvis Chen that concurrently show the mortality/morbidity rate by ABSM as well as the proportion of population in each socioeconomic stratum.
- Key changes since 2002 recommendations: Poverty data formerly collected as part of the 10-year decennial census are now collected annually as part of the American Community Survey from a smaller number of households requiring the use of data based on 3- or 5-year rolling averages.

EXAMPLES OF APPLICATIONS OF THIS APPROACH ON LOCAL/STATE PUBLIC HEALTH DATA

JIM HADLER, MD MPH _ Public Health Consultant, CT Emerging Infections Program, NYC Dept. of Health and Mental Hygiene

The only measure of socioeconomic status (SES) routinely used in the US is an indirect one, race/ethnicity. However, race/ethnicity is not very satisfying as an SES measure: it is not always available, it is getting harder to measure as there are increasing numbers of race/ethnic classifications, and disparities described by it are subject to misinterpretation and inappropriate generalizations. Neighborhood socioeconomic status (SES) can be used effectively to describe disparities in health status. However, it has not been used on a routine basis for

analysis or reporting of public health surveillance data, and there are no national standards for what neighborhood SES measure(s) should be used. Geocoding and linkage to ABSMs allows for the possibility of using SES measures.

Connecticut Emerging Infections Program (EIP)

- Objectives: The EIPs conduct gold-standard surveillance for a number of conditions agreed upon by the EIPs and CDC. A goal of the CT EIP is to gain experience using census tract neighborhood poverty level to describe disparities by routinely analyzing surveillance data using the PHDGP recommended method. Four surveillance system datasets were used, chosen based on pre-existing race-ethnic disparities: pneumococcal disease, influenza-associated hospitalizations, cervical cancer precursors, and food borne disease, beginning with campylobacter incidence. Results of analyses showed expected socioeconomic gradients, and also illuminated previously undetected socioeconomic disparities.
- Incidence of pneumococcal disease (IPD) by neighborhood used 3 poverty groupings vs. PHDGP recommended 4, and two time points for children, 1998-1999 and 2007-2008, before and after introduction of pneumococcal conjugate vaccine. Analysis in 1998-1999 found huge gradient overall and among whites (highest rates among poorest neighborhoods), but a slight reverse gradients among Hispanics. The overall SES gradient lessened by 2007-2008 for vaccine serotypes, but increased for non-vaccine types. While the SES gradient in whites followed the overall pattern for non-vaccine serotypes, the gradient in blacks and Hispanics did not, with biggest increased in the least impoverished neighborhoods,
- Incidence of influenza-associated hospitalizations in aggregate data of children under 18 years from 2003-2010 examined both census tract level poverty and crowding as measures of socioeconomic status, the latter using cut-points based on quartiles of percentage living in housing with at least one person per room. A similar and consistent SES gradient with rates highest in the most impoverished census tracts was found using each measure. The gradient was found for each individual year as well as all years combined.
- Surveillance for cervical cancer precursor analyses (CIN2, CIN3 and AIS) included pathology data from all sources. Because data lack race/ethnicity, geocoding allows researchers to look at disparities without needing race/ethnicity data. However, because a cervical cancer precursor diagnosis depends on having a Pap smear, results may be skewed because of screening issues. Overall rates among women 15 years and older had the expected SES gradient relative to poverty rates (lowest rates in wealthiest areas and highest in poorer areas). However, when examined by 5-year age group, a reverse SES gradient in 20-24 year olds was detected. When analyses were restricted to the major CT cities, the reverse gradient in the 20-24 age category basically went away. When analyses were restricted to suburban areas, the reverse gradient was pronounced -- likely due to increased screening.
- Food borne disease analysis -- lab confirmed campylobacter incidence -- examined 11 years worth of data. This revealed a reverse SES gradient by poverty level which persisted over time: more cases were seen in suburban/wealthier area. When examined by race-ethnicity, the reverse SES gradient was found for whites and Hispanics but not blacks. By age-group, it was found in all over 10 years but the gradient was reversed in those <10 years. This overall finding is consistent with case-control studies in which international travel and eating out in restaurants that serve chicken, both more likely with higher SES status, are risk factors for campylobacter. It also raises the question, though, whether the findings could reflect a higher rate of diagnostic workup (stool culture) for older children and adults living in wealthier neighborhoods.

New York City Department of Health and Mental Hygiene Workgroup

- Although the New York Department of Health and Mental Hygiene (DOHMH) did not have a standardized method for analyzing socioeconomic health disparities among NY programs, the DOHMH has been focusing on health disparities for a while. To address the need for a standard and standardized neighborhood

socioeconomic measure, a working group of DOHMH epidemiologists was commissioned to discuss possible standard measures for NYC, focusing on adapting previous work of the PHDGP.

- The noted challenges to employing the PHDGP methodology were (a) different population distribution from RI and MA (nearly 50% living in census tracts with >20% of the residents below the federal poverty level); (b) neighborhoods traditionally used by the DOHMH are 42 United Hospital Fund areas that are composites of ZIP codes; (c) a higher cost of living in NYC may make use of federal poverty level cutpoints less relevant. However, the workgroup agreed that the need for standard measures was important, and recommended that the background work done by the PHDGP research team be accepted, that the PHDGP methodology be routinely applied to DOHMH surveillance data, with some adjustment to the poverty cutpoints, i.e., utilization of 6 categories vs. 4 categories as necessary to reflect different population poverty distributions, particularly higher levels of census-tract level poverty.
- Analyses of all cause mortality and tuberculosis data demonstrate a strong SES gradient that persists through all 6 census tract poverty levels and within each of four major race-ethnic groups in NYC. For the comparison of rates both across poverty levels and across across race-ethnic groups, it was important to use age-standardized rates, particularly for mortality for which rates are age-dependent.

DISCUSSION OF MAJOR METHOD ELEMENTS FOR UPDATE/MODIFICATION/AGREEMENT

A. Geocoding Addresses

- PHDGP process: Utilized geocoding service.
- Krieger research team current practice: Geocoding performed in-house using ArcGIS
- For state/local/tribal epidemiologist: What are health departments currently doing re: geocoding? What are the take-home points? What shortcuts are available?

• DISCUSSION POINTS:

Agencies that routinely geocode data utilize different methods, e.g., the Georgia Department of Community Health uses Centrus (\$4000/year) and ArcGIS10 for special analysis, CDC uses ArcGIS, the Massachusetts Cancer Registry has a geocoder on staff who utilizes various methods. While there are a number of free web-based options for geocoding data, the issue of patient information confidentiality as pertains to submitting unencrypted addresses from health data to an outside source would need to be vetted by individual state/local/tribal health department IRBs.

The possibility of extending the capacity of 29 states currently funded by the CDC HIV program to geocode HIV data, to geocode data from state/local/tribal departments of health was explored, as was the possibility of funding an academic institution as the geocoding center (on behalf of CDC). This again brought up issues of patient data confidentiality, which could perhaps be addressed by formulating memorandums of understanding between the state academic partners and health departments and/or constructing a formalized data sharing agreement with the CDC.

Other issues regarding geocoding were also discussed, including: the need for familiarity with both the data and the local geography so that systematic inconsistencies and inaccuracies in the geocoding results can be detected; the need for a standard baseline percentage of completeness; whether or not to use post office box addresses; the extreme difficulty of geocoding in tribal areas where addresses are not standardized. The need for geocoding accuracy above all else was stressed, and it was generally agreed that the standard might not be the geocoding method, but rather the quality of the geocoding – which could be tested by using online sites such as the Census American FactFinder or GPS Visualizer to test a subset of 100 addresses per 100,000.

- **RECOMMENDATION:** Geocoded public health data have great value and can be used to link with area-based socioeconomic measures for monitoring health disparities. State and local public health agencies should consider geocoding some or all of their public health data for this use. While standardizing the geocoding method across health departments may not be feasible, a standard for geocoding accuracy is necessary.

B. Extracting Census Population Counts (denominators)

- PHDGP Process: Used 1990 US Census data.
- Krieger research team current practice: Awaiting release of the American Community Survey (ACS) 5-year population estimates on census tract level.
- For state/local/tribal epidemiologist: Which estimates to use for data up to 2010? And for data after 2010? What are health departments currently doing re: geocoding? What are the take-home points? What shortcuts are available?

- **DISCUSSION:**

The ACS population estimates are 5-year rolling estimates that can be used for this project for poverty estimates but perhaps not population estimates. The ACS surveys a sample of each state and projects the total population count, thus the estimate is only as good as the projection -- which is why the ACS recommends that they not be used for absolute population counts. The US 1990 and 2000 Census population counts are still available for analyses of health data around 1990 and 2000, and are fine for use as denominators. For analyses after 2005, we have the option of using the above-mentioned ACS estimates, however for 2010 and the years following, we can use the US 2010 decennial Census population counts, which should be released in the near future.

Additionally, once the 2010 census data are available, we can assume population increases linearly and interpolate the population for each census tract since the 2000 Census if we assume it increases at the same rate and/or by applying the proportions from the ACS to the population count.

- **RECOMMENDATION:** Use 1990, 2000, or 2010 decennial census tract data for population counts and explore using ACS data applied to annual state and county estimates later, when we've had more experience using them. For now, the simplest approach may be to use 1990 Census data for health outcomes from 1985-1994, 2000 Census data for health outcomes from 1995-2004, and 2010 Census data for health outcomes from 2005- 2014.

C. Constructing Area-Based Socioeconomic Measures, i.e., Poverty

- PHDGP Process: Used 1990 Census Data, % Poverty, with the cutpoints 0-4.9%, 5-9.9%, 10-19.9%, >20%
- Krieger research team current practice: Awaiting release of the American Community Survey (ACS) 5-year poverty estimates on census tract level.
- For state/local/tribal epidemiologist: Use census tract as base unit? What poverty cut points to use?

- **DISCUSSION:**

The PHDGP devoted considerable effort to showing the robustness of poverty as an area-based socioeconomic measure, and its use as a means of monitoring socioeconomic disparities in health that is readily understandable, i.e., as “% of persons living below the poverty level” vs. other commonly utilized measures e.g., “% living at 200% below poverty”. Yet there are other underlying factors that state/local/tribal health departments might also be interested in that warrant eventual consideration. Furthermore, as exemplified by the NYC and CT examples, universal cut points may not work for all groups, as some regions may require more

categories to illuminate granular data. 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 underserved area.

- **RECOMMENDATIONS:** For a useful census tract SES indicator, consider the poverty variable, “% of persons living at or below the federal poverty level.” Six poverty cutpoints (0-<5%, 5-<10%, 10-<20%, 20-<30%, 30-<40%, and >40%) should be considered for comparison across jurisdictions; these can be collapsed or expanded depending on the size of the population and poverty level spread within each jurisdiction. Independent of which cutpoints are used, we recommend allowing a break at <20%, >=20% to enhance comparability across jurisdictions and because at >=20% jurisdictions are eligible for some federal programs. Advanced users can also explore the use of other census tract poverty related indicators

D. Merging Dataset, Analyses

- PHDGP Process: Used SAS primarily
- Krieger research team current practice: SAS
- For state/local/tribal epidemiologist: How to show total poverty effects? And poverty effects by common demographic variables including gender, age groups, and race/ethnicity groups?

- **DISCUSSION:**

Although many software programs are currently used by the different state/local/tribal health departments, SAS is widely available and still used in most health departments, thus any programming made available to health departments should be distributed as a SAS template program that can be easily translated from SAS to other programs. Age-standardization should be performed where appropriate based on health outcome.

Race was generally not included in the PHDGP analyses due to incompleteness and/or unreliability in the health outcome datasets -- a common problem in health outcome data. Similarly, a lack of data on tribal affiliation, as well as a lack of specificity when categorizing individuals as simply “Asian”, are also current issues, as is the need to collect data in two emerging areas of study: disability and sexual identity.

The best use should be made of available race/ethnicity data, with additional effort expended to make the data more complete as opposed to excluding it. To this end, there was considerable discussion about imputing race data, citing examples of groups performing simple imputation by applying census tract race proportions to missing data, as well as groups following the literature that recommends multiple imputations up to a maximum of 40% missing data. Noting that this is a step that can get extremely complicated and resources to address this issue may not be available at all health departments, this may be a step that is not appropriate to come to agreement at the outset of this project.

- **RECOMMENDATION:** There are several statistical softwares that can be used to merge and analyze datasets. SAS has been popular among several users. SAS code, along the lines of the code presented on the PHDGP website, should be made available to CSTE member organizations to facilitate performing these analyses. The decision to age-standardize rates should be based on health department standard practices vis a vis specific health outcomes. At this point, there is no recommendation for data imputation for missing data. Where race data is of sufficient quality to include in analyses, use OMB categories to the extent possible, following NCHS guidelines about data stability and limits of data suppression.

E. Showing Output

- PHDGP Process: Histograms, scatterplots, choropleth maps (mortality data only)

- Krieger research team current practice: Histograms, with dual representation of magnitude of rate and percent of affected population, as developed by Dr. Jarvis Chen
- For state/local/tribal epidemiologist: Which are simplest to produce, to understand, and to compare across jurisdictions?

- **DISCUSSION:**

Whatever form of output is recommended, it needs to be understandable to all audiences and as user-friendly as possible. Several examples of useful output formats were discussed, including: the PHDGP histograms; simple histograms with accompanying text regarding the proportion of population represented by each bar; NCHS histograms using dual y-axes; and bubble graphs. The new version of EpiInfo will have the capacity to output weighted bar graphs with confidence intervals, akin to the PHDGP weighted bar graphs, and will be able to pull data from SQL server and excel spreadsheets.

Mapping would not be the best initial standard of output for this project, because of confidentiality issues and the instability of rates on such a small, i.e., census tract, level.

- **RECOMMENDATION: Simple histograms, with accompanying text, should be the first standard form of output for consideration. Additional text guiding interpretation of the data should routinely be included on all output.**

OTHER CONSIDERATIONS

- Guidance format: A web-based visualization for the guidance is easier than a printed guidance.
- Missing data: In general, using multiple imputations to address missing data – whether missing post office box addresses or race/ethnicity – will be addressed at a later stage of the project.
- Interpretation of data: As noted above, text that guides interpretation of the concept presented, e.g., that distinguishes between the characteristics of the census tracts vs. the characteristics of the individuals who reside therein should be routinely included in all output. This text will be supplied by the PHDGP.
- Tribal Health: Tribal health departments may face challenges in implementing this methodology, but may also benefit from less strenuous concerns about confidentiality when sharing data within tribe, as opposed to outside.

SELECTING A NATIONALLY NOTIFIABLE DISEASE OR CONDITION FOR DEMONSTRATION

- PHDGP Process: Studied 7 health outcomes: (mortality (all cause and cause-specific), low birthweight, childhood lead poisoning, sexually transmitted infections, tuberculosis, and non-fatal weapons-related injuries)
- Krieger research team current practice: Studying a variety of health outcomes
- For state/local/tribal epidemiologist: What health outcome to apply this approach to?

- **DISCUSSION:**

A nationally notifiable disease is preferred because all states collect that data, but a health outcome with a well-known race disparity might also be a good choice to showcase the methodology. STIs, with an emphasis on gonorrhea and/or chlamydia, and HIV were discussed, but issues with missing race/ethnicity data will be a

CSTE Health Disparity Methodology Meeting, Atlanta _ May 24, 2011
Summary

factor. Additionally, the CDC is already geocoding HIV data in 29 states. Launching the CSTE project focusing on the same health outcome might limit new involvement.

TB and/or hepatitis, teen births, obesity, cardiovascular diseases, tobacco control, seasonal influenza vaccination, cancer, diabetes mortality, and mortality data were all discussed as possible choices for the demonstration health outcome. Infant mortality was also suggested as the simplest health outcome to begin, as it would not require the step of age-standardization.

Mortality data (recommended by the PHDGP) as a starting point for the project have the strength of being a particularly robust dataset, i.e., large numbers and fairly complete data. All cause mortality might also be a good choice for the demonstration because it cuts across all diseases. If a specific mortality cause were chosen, there may be difficulty agreeing on one. It was noted that mortality data goes from states to NCHS to CDC without identifiers, so the CDC would need to request approval to get addresses and share the data with CSTE members. Individual CSTE member health departments with access to mortality data should be able to use their state's data.

- **RECOMMENDATION:** There will be further discussion on which nationally notifiable diseases to be considered for a demonstration project.