Lessons Learned From the Environmental Public Health Tracking Sub-County Data Pilot Project

Angela K. Werner, PhD, MPH, **Heather Strosnider, MPH**, **Craig Kassinger, BS**, **Mikyong Shin, DrPH, MPH, RN**, and **Sub-County Data Project Workgroup**

Division of Environmental Hazards and Health Effects (Drs Werner and Shin, Ms Strosnider, and Mr Kassinger) and ORISE Postdoctoral Fellow, Environmental Public Health Tracking Branch (Dr Werner), National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgia.

Abstract

Objective: Small area data are key to better understanding the complex relationships between environmental health, health outcomes, and risk factors at a local level. In 2014, the Centers for Disease Control and Prevention's National Environmental Public Health Tracking Program (Tracking Program) conducted the Sub-County Data Pilot Project with grantees to consider integration of sub-county data into the National Environmental Public Health Tracking Network (Tracking Network).

Design: The Tracking Program and grantees developed sub-county-level data for several data sets during this pilot project, working to standardize processes for submitting data and creating required geographies. Grantees documented challenges they encountered during the pilot project and documented decisions.

Results: This article covers the challenges revealed during the project. It includes insights into geocoding, aggregation, population estimates, and data stability and provides recommendations for moving forward.

Conclusion: National standards for generating, analyzing, and sharing sub-county data should be established to build a system of sub-county data that allow for comparison of outcomes, geographies, and time. Increasing the availability and accessibility of small area data will not only enhance the Tracking Network's capabilities but also contribute to an improved understanding of environmental health and informed decision making at a local level.

Correspondence: Mikyong Shin, DrPH, MPH, RN, Division of Environmental Hazards and Health Effects, National Center for Environmental Health, Centers for Disease Control and Prevention, 4770 Buford Hwy NE, MS F60, Atlanta, GA 30341 (mshin@cdc.gov).

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

The authors declare no conflicts of interest.

Keywords

census tract; environmental health; lessons; small area; sub-county; surveillance; tracking

Small area data are essential to better understand complex environmental health problems at a local level, with higher-resolution data necessary for uncovering local rate variation. Smaller areas, unlike county level, generally preserve the geographic variability of data.[1] Small area analyses can increase our knowledge of local health issues. They can help inform public policy, have the potential to serve as surveillance tools, identify geographic areas where care can be improved, and provide information for creating and targeting specific interventions.[2,3] Small area data can also be used to help attribute causality to observed health outcomes.[4]

The use of spatial analysis methods has grown rapidly in recent years,[2] as has the use of small area data. This growth has promoted increased availability and accessibility of these data, as well as an increasing demand to examine relationships between the environment, risk factors, and health outcomes in local areas. This has prompted the Centers for Disease Control and Prevention's (CDC's) National Environmental Public Health Tracking Program (Tracking Program) to consider building a system of sub-county data within the National Environmental Public Health Tracking Network (Tracking Network).

The Tracking Program was established in 2002 to track exposures and health effects associated with environmental hazards and to bridge existing data gaps.[5] Generally, fewer data are available at the sub-county level.[6,7] The current geographic resolution of data collected by the Tracking Program is limited mostly to the county or state level. The Tracking Program is working to add more sub-county data to fill this void, creating a system of sub-county data. Considerations specific to the Tracking Program's decision-making process include temporality (eg, cross-sectional vs longitudinal), compatibility between data and measures, dramatic increases in data that must be managed, and communications and technical issues with displaying data.

To address this gap, the Tracking Program had to identify the challenges of increasing the availability of standardized sub-county data and propose solutions to overcome these challenges. A pilot project was launched to shed light on these issues and gain insights into future work at the sub-county level. State and local Tracking Programs were to develop or enhance standards as part of the Nationally Consistent Data and Measures (NCDM) for sub-county data calculation, dissemination, and display. Three overarching questions were identified:

1. How can sub-county data be displayed on the Tracking Network to ensure maximum utility and protection of confidentiality?

2. What are the barriers to accessing and sharing sub-county data through the Tracking Network?

3. What are the challenges to implementing data submission standards and display of sub-county data on the Tracking Network?

Increasing the availability of standardized sub-county data is important for enhancing the capability of the Tracking Network, improving our understanding of environmental health, and informing local-level decision making. This article describes the project and then discusses the lessons learned and recommendations to increase the availability of sub-county data within the Tracking Network and to guide others working with similar data.

Project Overview

In 2014, the Tracking Program proposed the Sub-County Data Pilot Project. Tracking Program grantees were invited to apply for funding to work on this topic. Florida, Maine, New York, Washington, and Wisconsin were selected to develop sub-county-level data from August 1, 2014, to July 31, 2015. Through individual and monthly conference calls, the team evaluated the availability of sub-county data within each state and developed data standards. The team heard expert presentations on the Geographic Aggregation Tool (GAT),[8] developed by the New York State Department of Health, and calculations for small area data. At the end of the project, grantees submitted sub-county data to CDC. The Figure presents an overview of the process.

The team reviewed available sub-county data of interest to each grantee. Two data sets were selected for which the Tracking Program already had corresponding county-level NCDM and which were available to 4 of the 5 grantees. These included data sets on acute myocardial infarction (AMI) from hospitalization data and low birth weight (LBW) from birth certificate data. Maine selected childhood lead and private well water data. Grantees submitted data at varying geographic levels, depending on availability.

Existing NCDM standards were extended by developing new standards for geographic and temporal aggregation, suppression, and calculation methods. The team developed standardized indicator measures, data dictionaries, and how-to guides, based on geographic and temporal aggregation levels and state suppression policies for data at set geographic levels. This allowed for standardized extraction of required data elements, identification of cases and events, and calculation of measures from sub-county data. Because of inconsistencies in available geographies, a standard geography was not created. The team did standardize the processes for creating geographies and submitting data.

Each health data set had a variable to uniquely identify sub-county units, which were linkable to geographic data developed by grantees. Geographic data were based on individual or aggregated geographies by census tract, zip code, or town. The Table details the data submission options for each grantee. Data were submitted by all grantees, resulting in a successful data call. Grantees were instructed to document their decisions and challenges in working with sub-county data over time to better inform future sub-county data use and processes within the Tracking Program and more broadly. At the end of the project, CDC and the grantees reviewed lessons learned and recommendations resulting from this work.

Lessons Learned

The needs of state, local, and community-based programs need to be weighed against the challenges of using finer resolution data, which is a trade-off between geographic resolution and statistical stability.[9] Small area data are more challenging to work with than are larger-scale data because of changing boundaries, lack of historical patterns of change to serve as a basis for estimation, reliability issues, and location-specific factors that can greatly affect calculations.[10,11] The team encountered many of these issues during the pilot project. Several main challenges emerged throughout the process, including during geocoding and assessing aggregation needs. Challenges identified as "lessons learned" included inconsistent geographic levels, aggregation, geocoding data, and data stability issues.

Available geographic level

For this project, no one data set was available for which all grantees had the same geographic resolution. Census tract, town, and zip code were used for different data sets. Zip codes are identifiable for the public and available in many data sets. However, using zip codes has disadvantages.[12–15] Zip code boundaries are spatially and temporally dynamic, differing from other geographies in terms of stability.[14,15] Many grantee data sets were only available by zip code, which must be converted to zip code tabulation areas (ZCTAs). ZCTAs are used to address the difficulties in defining areas covered by zip codes and to assign population estimates.[15,16] However, ZCTAs do not match zip codes exactly and the areas they cover might differ, so typical public health surveillance data cannot be easily linked to ZCTAs.[12,15,17] The notion that ZCTAs represent finer levels of geographic resolution is certainly not the case for exurban and rural areas.[14,18]

Census tracts belong to a hierarchical structure of spatial units created by the US Census Bureau. In comparison with zip code areas and ZCTAs, census tracts change less frequently, allowing for longitudinal analyses.[13,19,20] Krieger et al[21] have promoted the use of census tract–level data due to the relatively homogenous populations and that they comprise administrative units used by various agencies. Some studies concluded that census tract (or block group) units should be used, as these performed better when detecting socioeconomic gradients across various health outcomes than zip codes.[12,22] Census tracts, however, are a less recognizable unit when displaying data. In addition, census tracts are considered coarse spatial units when aggregating certain health outcomes (eg, cancer) and estimating exposures.[23] Geocoding to this level is often based on billing addresses, which might not reflect place of residence. Some data sets, such as vital records, are geocoded using residential address.[24]

Creation of a national standard geography would, in some ways, reduce the time, effort, and resources needed to ensure consistency across the Tracking Network. The selection of sub-county geographies should provide a compromise between confidentiality issues and the demand for finer geographic resolution data.[18] However, data availability at this resolution has limitations. The selection of a set of geographies ultimately should allow for comparison across outcomes, over time, and between places. For monitoring purposes, dissimilarities of area-based measures across various geographic levels hinder their use for comparisons.[12] The use of such different geographies (eg, census tract, town, zip code) does not allow for

comparability and also creates display issues and inconsistency for display on the Tracking Network. Hence, the need for uniform measures and geographies.

Aggregation

Most grantees found that sub-county data were too sparse to present stable, unsuppressed rates. Temporal aggregation can be used, in which a certain number of years are combined for a given area. Grantees noted the need for aggregation over a 3- or 5-year period, depending on the data set. A 5-year period was used for standardization. Temporal aggregation is generally used to estimate stable rates over time, particularly with rare outcomes. It is also used to ensure confidentiality.[25] However, any estimates relying on data aggregated temporally will not be able to show trend differences over time for smaller areas. [26] The suitability and necessity for using temporal aggregation will be highly dependent on the data set.

Oftentimes, geographic (spatial) aggregation is needed both to obtain meaningful units for analysis[9] and to protect confidentiality.[25,27] Several grantees used the GAT[8] to facilitate this, which gave the stability needed to monitor outcomes over time and for grantees to generate the best geography for temporal data points. However, visualization and analysis of long-term trends can be difficult with inconsistent geographies. Grantees noted that aggregations should be meaningful to the public, if possible (eg, zip code aggregation eliminates the ability to recognize the geography). When using geographic aggregation, numerator and denominator rules should be established to have cut points for aggregation or suppression. If, for example, areas are too small, spurious spatial patterns from random variation are likely. [28,29]

Although geographic aggregation is often necessary, it presents the modifiable areal unit problem (MAUP), where "conclusions based on data aggregated to a particular set of districts may change if one aggregates the same underlying data to a different set of districts."[9(p104)] The scale (aggregation) effect occurs when the same data are grouped into larger areal units, resulting in different effects. The grouping (zoning) effect occurs when results differ because of alternative areal unit formations at similar scale.[30] MAUP cannot be solved; however, it is a problem that should be recognized, understood, and planned for when working with spatial data and geospatial analyses.[27,31]

For either form of aggregation, the minimum percentage of data that can be suppressed needs to be established. This could be based either on the percentage of geographies (eg, percentage of census tract aggregation suppressed) or on the percentage of total cases (eg, percentage of hospitalizations out of the total number of cases for that outcome across all geographies).

Geocoding

Geocoding occurs when a spatial location is assigned to an address record.[9] Thus, records are matched in at least 2 databases: one with address information and one with a reference file containing addresses and geographic coordinates.[9] Multiple errors can arise with geocoding, including records that have incorrectly recorded addresses or records that have correctly recorded addresses but incorrect geocodes.[32] While it would be ideal to have point

data to improve our analyses of potential environmental exposures at the sub-county level, many institutions do not share these data or have access restriction policies to ensure confidentiality.[9] The Tracking Program will be geocoding to census tracts rather than relying on point data. In turn, this means that studies (and the conclusions drawn from those studies) can be limited.[9]

The team suggested establishing a set of guidelines for geocoding data, particularly for dealing with records that are more difficult to geocode. This will allow for more accurate, consistent geocoding. Some grantees suggested exploring alternative reference files so that the number of accurately geocoded records can be increased, reducing the need for manual geocoding. Geocoding to the zip code level should be avoided, if possible, as exact spatial boundaries are often unknown, and frequent boundary changes mean that a geocoding file could be outdated.[33]

Nongeocoded records would not be mapped and aggregated. If these data are ignored, this could result in bias toward urban areas, resulting in misleading rates. To deal with missing geography, some grantees used imputation to assign census tracts to nongeocoded records. For example, a zip code contained within a census tract could be assigned to that tract. When a point zip code (eg, for a high-volume address or PO box) is given, an enclosing zip code for the area should be assigned and assessed to see whether it is within a census tract. If it is not, it should be imputed using race or age.

Issues also arose with geocoding geographic units that had a population of zero or that were listed as a PO box. Where zip codes must be geocoded, the team suggested geocoding addresses to the zip code centroid. This is typically done when using commercially available zip code boundary approximations.[33] It is also important to know which, if any, zip codes enclose a PO box–only zip code (ie, a "P" classification zip code).

Population data–related challenges

Accurate population estimates are crucial for accurate rate calculation. However, using data from the census, conducted every 10 years, is not sufficient for these purposes, so estimation methods are required.[34] Available sub-county population estimate data were discussed during this project to determine which were most suitable. Estimates are available from the American Community Survey, state and local government agencies, and private organizations.[35] However, certain estimates worked better for some grantees than others. Therefore, different estimates and estimation methods were used for this pilot project. Key factors to consider when selecting population estimation methods are the quality and type of data available.[35] Several criteria, including estimation error and uncertainty, necessary detail, validity, plausibility, cost, timeliness, and ease of application, should be used to evaluate estimates.[36]

Other population-related data issues were identified during this project, including the following:

- Inaccurate or misleading data, such as significant census population miscounts in some areas;

- Challenges with the underlying population (eg, prisons, universities); and

- Geographies that had cases but no populations.

Some states have large areas of federal lands that cannot logically be aggregated into nearby areas; therefore, the map has to accurately reflect no populations or rates for these areas. Rather than assign zero population to these areas, it is better to classify these as no measurement so that calculations are not distorted.[37] Some patients were also double-counted in hospital transfers. This happened more often in rural areas than in urban areas. Double-counting of patients can be problematic, possibly introducing bias into a data set.[38]

While the geographies discussed earlier focus on administratively defined units, another consideration for future work is the use of gridded population data. This would eliminate the need for determining which geographies grantees have available and allow for normalizing geographic space, addressing the demand for up-to-date population data at a higher resolution, and obtaining population and environmental data at mutually compatible scales. [39] There are currently several high-resolution gridded population data sets that have been used in epidemiological studies (Gridded Population of the World, Global Rural-Urban Mapping Project, LandScan)[40]; however, there are differences between these databases and between these databases and reference data.[41] For example, LandScan effectively handled gridded areas where no one lives while the others did not.[41] Such data sets would have to be examined to determine whether they are suitable for Tracking's needs.

Data stability and other data challenges

Data stability is an important issue. Data aggregation using smaller geographic units might be more accurate, but less stable rates can arise because the area includes fewer at-risk persons.[27] This can also occur when using rarer health outcomes at finer resolutions.[42] Data stability is of particular concern when rates are based on small numbers, which is likelier in rural areas. In these cases, suppression and aggregation rules do not necessarily guarantee rate stability and rates can vary widely by chance. The team also found lower rates near state borders. Hospitals near state borders often treat patients from neighboring states,[43] which requires neighboring states' data to ensure accurate calculations.

The standard approach to calculating 95% confidence intervals (CIs) can perform poorly with small area data due to non-normal data distribution. The reliability of estimates can also be assessed using relative standard error (RSE), a measure indicating the extent to which estimates deviate from true values, expressed as a fraction of the estimate.[44] Typically, an RSE greater than 25% indicates high sampling error.[44] For this project, alternatives considered for calculating CIs included ABC intervals and Dobson, Kuulasmaa, Eberle, and Scherer intervals.[45] A third alternative, the adjusted gamma CI estimator, was used for rates based on small numbers.

The potential for rate variation means there can be misinterpretation in mapping displays. This should be considered when data are shared. Data visualization, such as the Tracking Network's maps on the Internet, is a means of sharing information, getting the viewer's attention, and attracting interest in the data.[28,46] Specific to small area analyses, rates that are mapped in smaller areas are often somewhat misleading because of unstable rates, which

is why smoothing techniques are often preferred in these scenarios.[28] Other visualization considerations include choosing appropriate administrative units, data classification methods, and color schemes or hatching patterns.[28]

In cases where risk estimates are unstable, Bayesian hierarchical models have been used to address sparseness in populations and cases, allowing for an adaptive smoothing approach.[47] This can, however, create overly smoothed maps, masking true risk distribution.[47] The degree of smoothing used is a trade-off between high sensitivity and high specificity.[29] Some have suggested that a numerator of 20 or more is needed to produce fairly stable estimates, approximating a normal distribution and allowing for simpler CI calculations.[48] There are also other analyses and modeling approaches to consider, including spatial regression techniques. One such example is geographically weighted regression, which allows for regression coefficients to vary over space.[49] An underlying principle of this approach is that it is expected that places that are closer together are more similar than those greater apart (ie, they are spatially autocorrelated).[50] Approaches such as this would allow the Tracking Program to explore, analyze, and model relationships between covariates and outcomes of interest while accounting for spatial variation.

For maps, uncertainty might need to be displayed to provide a clearer picture of the data, which is not unique to small area data but important to consider. Several options can be used to display these uncertainties alongside risk estimates.[47] Estimates and uncertainties can be displayed on a bivariate choropleth map,[51] for example. They can be shown using opacity to represent uncertainty,[52] decreasing boundary crispness where more uncertainty exists, or using posterior probability values for the interpretation of areas with excess risk.[47,53] Ideally, a standard method for displaying and communicating these uncertainties should be used to ensure consistency.

Another challenge noted during this project was the lack of consistent state and national standards for analyzing, sharing, and displaying sub-county data. Therefore, for this project, different outcome data (eg, AMI, LBW, childhood lead), geographic boundaries (eg, zip code, census tract), and denominator data were used. Establishing standards would improve the maintainability and scalability of data on the Tracking Network's public portal and increase the comparability between data sets to allow for comparison across studies, time, and places.

Recommendations

The Tracking Program undertook this project to increase the availability and accessibility of sub-county data on the Tracking Network while considering the Tracking Program's unique needs for creating a system of sub-county geography over time. The considerations and lessons learned from this project led to several key recommendations. These are important steps for building this system of sub-county data and using the data in a way that provides meaningful analyses, particularly longitudinal analyses.

First, data stewards should be engaged to develop new data sharing agreements and policies to allow for the analysis and dissemination of sub-county data for surveillance purposes.

Collaboration with data stewards will also include efforts to geocode address-level data to census tracts to allow for standardization. Tracking Program and grantees need guidelines for how to handle difficult records and missing geographies so that all use accurate, consistent methods. Alternative reference files also could be considered, which might reduce the time and resources needed for manual geocoding. Perhaps, using the same software or company for geocoding might ensure consistency among grantees.

Second, the Tracking Program and grantees, in collaboration with data stewards and data users, should develop standardized sub-county geographies. This would be done by aggregating census tracts to allow for comparison and tracking of data over time while meeting the needs of data users. This might include a more conservative aggregation scheme for rarer outcomes (eg, birth defects) and a less conservative aggregation scheme for more common outcomes (eg, emergency department visits for asthma). These schemes should balance the need for sub-county geographies and stable rates using minimal suppression. Potential solutions should be evaluated for identified issues, including handling of census tracts with zero population and zero cases or with zero population and cases, managing and validating census tract–level data, and recognizing and planning for the MAUP. Methods are required both to deal with changes to census tracts over time (ie, creating a consistent geography) and to understand how proposed aggregations could be affected by these changing boundaries.

Third, available population estimates and estimation methods should be evaluated to identify the best ones for these purposes. Guidelines should be established to direct how to incorporate estimate uncertainties into analyses and displays. Appropriate methods should be outlined for calculating sub-county rates, RSE, and CIs. Guidelines also should be established for dealing with reliability issues posed by small numbers frequently encountered in sub-county data sets. Associated uncertainties should be factored into rate calculations and data displays.

Finally, collaboration is key to advancing these efforts and making sub-county data available on a larger scale. The Tracking Network and grantees cannot undertake this endeavor alone. Data stewards need to be involved to work toward the goal of increasing the availability and accessibility of sub-county data. Other experts should be engaged to help develop standards and evaluate the selected methods. In addition, training is required for all aspects of sub-county data use and analysis to ensure successful integration, analysis, dissemination, interpretation, and communication of sub-county data.

The use of sub-county data can also increase public awareness of place-based factors and understanding how these relate to health. Examining socioeconomic disparities at this finer geographic resolution can help guide resource allocation and set and evaluate health objectives.[21] Having an available system of sub-county data through the Tracking Network will allow users to better understand local health outcomes and risk factors over time. Incorporating these data into the Tracking Network requires collaboration with data stewards and adequate training of public health practitioners so that the benefits of using these data can be fully realized and identified challenges resolved.

Acknowledgments

References

1. Talbot TO, Haley VB, Dimmick WF, Paulu C, Talbott EO, Rager J. Developing consistent data and methods to measure the public health impacts of ambient air quality for Environmental Public Health Tracking: progress to date and future directions. Air Qual Atmos Health. 2009;2(4):199–206. [PubMed: 20098503]

2. Auchincloss AH, Gebreab SY, Mair C, Diez Roux AV. A review of spatial methods in epidemiology, 2000–2010. Annu Rev Public Health. 2012;33(1):107–122. [PubMed: 22429160]

3. Smith SK. Small-area analysis In: Demeny P, McNicoll G, eds. Encyclopedia of Population. New York, NY: Macmillan Reference USA; 2003:898–901.

4. Fawcett S, Holt C, Schultz J, Rabinowitz P. Section 22: using small area analysis to uncover disparities. In: Chapter 3: Assessing Community Needs and Resources. http://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/small-area-analysis/main. Published 2016. Accessed May 8, 2017.

5. McGeehin MA, Qualters JR, Niskar AS. National Environmental Public Health Tracking Program: bridging the information gap. Environ Health Perspect. 2004;112(14):1409–1413. [PubMed: 15471734]

6. Bemis K, Gray S, Patel MT, Christiansen D. Disproportionate emergency room use as an indicator of community health. Online J Public Health Inform. 2016;8(1). http://ojphi.org/ojs/index.php/ojphi/article/view/6409. Accessed December 21, 2016.

7. US Centers for Disease Control and Prevention. Community Health Assessment for Population Health Improvement: Resource of Most Frequently Recommended Health Outcomes and Determinants. Atlanta, GA: Office of Surveillance, Epidemiology, and Laboratory Services; 2013.

8. Talbot TO, LaSelva GD. Geographic Aggregation Tool, version 1.31. Troy, NY: New York State Health Department; 2010.

9. Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Hoboken, NJ: John Wiley & Sons Inc; 2004.

10. Murdock SH, Cline M, Zey M. Challenges in the analysis of rural populations in the United States In: Kulcsar LJ, Curtis KJ, eds. International Handbook of Rural Demography. Dordrecht, the Netherlands: Springer; 2012:7–15.

11. Rayer S Demographic techniques: small-area estimates and projections In: Wright JD ed. International Encyclopedia of the Social & Behavioral Sciences. Oxford: Elsevier Science Ltd; 2015.

12. Krieger N, Chen J, Waterman P, Soobader M, Subramanian S, 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(3):186–199. [PubMed: 12594195]

13. Grubesic TH, Matisziw TC. On the use of zip codes and zip code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. Int J Health Geogr. 2006;5(1):58. [PubMed: 17166283]

14. Grubesic TH. Zip codes and spatial analysis: problems and prospects. Socio Econ Plann Sci. 2008;42(2):129–149.

15. 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(7):1100–1102. [PubMed: 12084688]

16. US Census Bureau. Zip codeTM tabulation areas (ZCTAs™). https://www.census.gov/geo/reference/zctas.html. Published 2015. Accessed January 5, 2017.

17. US Census Bureau. Frequently asked questions: ZCTAs. https://ask.census.gov/prweb/PRServletCustom/YACFBFye-rFIz_FoGtyvDRUGg1Uzu5Mn*/!STANDARD#. Published 2015. Accessed January 5, 2017.

18. Johnson GD. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. Int J Health Geogr. 2004;3(1):29. [PubMed: 15588279]

19. US Census Bureau. Geographic terms and concepts—geographic presentation of data. https://www.census.gov/geo/reference/gtc/gtc_geopres.html. Published 2015. Accessed January 5, 2017.

20. US Census Bureau. Geographic terms and concepts—census tract. https://www.census.gov/geo/reference/gtc/gtc_ct.html. Published 2012. Accessed January 5, 2017.

21. 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(2):312–323. [PubMed: 15671470]

22. 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(5): 471–482. [PubMed: 12196317]

23. Jacquez GM. Current practices in the spatial analysis of cancer: flies in the ointment. Int J Health Geogr. 2004;3(1):22. [PubMed: 15479473]

24. Fitzgerald E, Wartenberg D, Thompson WD, Houston A. Birth and fetal death records and environmental exposures: promising data elements for environmental public health tracking of reproductive outcomes. Public Health Rep. 2009;124(6):825–830. [PubMed: 19894425]

25. Bell BS, Hoskins RE, Pickle LW, Wartenberg D. Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. Int J Health Geogr. 2006; 5(1):49. [PubMed: 17092353]

26. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. Am J Prev Med. 2004;26(5):453–460. [PubMed: 15165663]

27. Nelson JK, Brewer CA. Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. Cartogr Geogr Inf Sci. 2017;44(1):35–50.

28. Rezaeian M, Dunn G, St Leger S, Appleby L. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. J Epidemiol Community Health. 2007;61(2):98–102. [PubMed: 17234866]

29. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. Environ Health Perspect. 2004;112(9):998–1006. [PubMed: 15198920]

30. Gotway CA, Young LJ. Combining incompatible spatial data. J Am Stat Assoc. 2002;97(458):632–648.

31. Butkiewicz T, Meentemeyer RK, Shoemaker DA, Chang R, Wartell Z, Ribarsky W. Alleviating the modifiable areal unit problem within probe-based geospatial analyses. Comput Graph Forum. 2010;29(3):923–932.

32. 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(7): 1114–1116. [PubMed: 11441740]

33. California Environmental Health Tracking Program. Frequently asked questions about geocoding. http://cehtp.org/faq/tools/frequently_asked_questions_about_geocoding. Published 2015. Accessed January 5, 2017.

34. Swanson DA, McKibben JN. New directions in the development of population estimates in the United States? Popul Res Policy Rev. 2010;29(6):797–818. [PubMed: 21076679]

35. Bryan T Population estimates In: Siegel JS, Swanson DA, eds. The Methods and Materials of Demography. 2nd ed. San Diego, CA: Elsevier Academic Press; 2004:523–560.

36. Swanson DA, Tayman J. Subnational Population Estimates. Dordrecht, the Netherlands: Springer; 2012.

37. Harris R Quantitative Geography: The Basics. London, England: Sage Publ Ltd; 2016.

38. Kristoffersen DT, Helgeland J, Clench-Aas J, Laake P, Veierød MB. Comparing hospital mortality—how to count does matter for patients hospitalized for acute myocardial infarction (AMI), stroke and hip fracture. BMC Health Serv Res. 2012;12(1):364. [PubMed: 23088745]

39. Diechmann U, Balk D, Yetman G. Transforming Population Data for Interdisciplinary Usages: From Census to Grid Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), CIESIN, Columbia University.

40. Galway LP, Bell N, SAE AS, et al. A two-stage cluster sampling method using gridded population data, a GIS, and Google EarthTM imagery in a population-based mortality survey in Iraq. Int J Health Geogr. 2012;11(1):12. [PubMed: 22540266]

41. Hall O, Stroh E, Paya F. From census to grids: comparing gridded population of the world with Swedish census records. Open Geogr J. 2012;5:1–5.

42. Jarup L, Best N. Editorial comment on geographical differences in cancer incidence in the Belgian Province of Limburg by Bruntinx and colleagues. Eur J Cancer. 2003;39(14):1973–1975. [PubMed: 12957450]

43. Agency for Healthcare Research and Quality. Introduction to the HCUP Nationwide Readmissions Database (NRD) 2014. Silver Spring, MD: Agency for Healthcare Research and Quality; 2016.

44. Australian Bureau of Statistics. What is a standard error and relative standard error? Reliability of estimate for labour force data. http://www.abs.gov.au/websitedbs/d3310114.nsf/Home/What+is+a+Standard+Error+and+Relative+Standard+Error,+Reliability+of+estimates+for+Labour+Force+data#Anchor3. Published 2010. Accessed January 10, 2017.

45. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Stat Med. 1997;16:791–801. [PubMed: 9131766]

46. Everitt BSE, Dunn G. Applied Multivariate Data Analysis. London, England: Arnold; 2001.

47. Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. Environ Health Perspect. 2008;116:1105–1110. [PubMed: 18709139]

48. Shah GH. A Guide to Designating Geographic Areas for Small Area Analysis in Public Health: Using Utah's Example. Salt Lake City, UT: NAHDO-CDC Cooperative Agreement Project; 2005.

49. Wheeler DC. Geographically weighted regression In: Fischer MM, Nijkamp P, eds. Handbook of Regional Science. Berlin, Germany: Springer; 2014:1435–1459.

50. Lloyd CD. Spatial Data Analysis: An Introduction for GIS Users. Oxford, England: Oxford University Press; 2010.

51. Monmonier M Cartography: Uncertainty, interventions, and dynamic display. Prog Hum Geogr. 2006;30(3):373–381.

52. Drecki I Visualization of uncertainty in geographical data In: Fisher PFG, Goodchild MF, eds. Spatial Data Quality. London: Taylor & Francis; 2002:140–159.

53. Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect. 2004;112(9):1016–1025. [PubMed: 15198922]

Author Manuscript

Implications for Policy & Practice

Future use of sub-county data can have important implications for public health. It can support

- Identification and monitoring of health disparity hotspots;

- Investigation of contributing behavioral, social, and environmental factors; and

- Examination of health outcome variations across time, populations, and places.
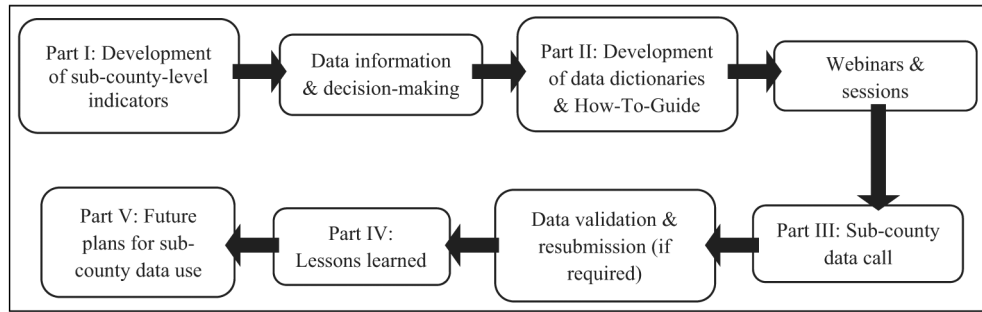
FIGURE.
Overview of the Sub-County Data Pilot Project Process

TABLE

Overview of the Data Submitted by Each Grantee for the Pilot Project

	Florida	Maine	New York State	Washington	Wisconsin
Geography level	CT: LBW Zip code: AMI	HMP: CL HMP and town: PWW	CT: AMI and LBW	CT: LBW Zip code: AMI	Zip code: AMI and LBW
Need for geographic aggregation	No	N/A[a]	Yes (using GAT)	Either geographic or temporal (eg, single year with CT aggregation)	Yes (using GAT)
Need for temporal aggregation	Yes (5 y, 2009–2013)	Yes (5 y for CL, 2009–2013)	Yes (5 y, 2008–2012)	Yes (5 y, 2009–2013)	Yes (5 y, 2009–2013)
State suppression policy	None	Suppress counts <6	No sub-county-level data sharing (must submit rates)	Small numbers guidelines[b]	Suppress counts <5

Abbreviations: AMI, acute myocardial infarction; CL, childhood lead; CT, census tract; GAT, Geographic Aggregation Tool; HMP, Healthy Maine Partnership service areas; LBW, low birth weight; N/A, not applicable; PWW, private well water.

[a] HMPs are aggregations of towns.

[b] Washington State Department of Health Guidelines for Working with Small Numbers can be accessed at http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf