HHS Public Access

Author manuscript

Spat Spatiotemporal Epidemiol. Author manuscript; available in PMC 2022 February 14.

Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2020 June ; 33: 100339. doi:10.1016/j.sste.2020.100339.

Developing a surveillance system of sub-county data: Finding suitable population thresholds for geographic aggregations

Angela K Werner[a,b,*], **Heather M Strosnider**[a]

[a]Division of Environmental Health Science and Practice, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA, United States

[b]ORISE Postdoctoral Fellow at the Environmental Public Health Tracking Section, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Abstract

The Centers for Disease Control and Prevention's National Environmental Public Health Tracking Program created standardized sub-county geographies that are comparable over time, place, and outcomes. Expected census tract-level counts were calculated for asthma emergency department visits and lung cancer. Census tracts were aggregated for various total population and sub-population thresholds, then suppression and stability were examined. A total of 5,000 persons was recommended for the more common outcome scheme and a total of 20,000 persons was recommended for the rare outcome scheme. Health outcomes with a median case count of 17.0 cases or higher should produce stable estimates at the census tract level. This project generated recommendations for three sub-county geographies that will be useful for surveillance purposes: census tract, a more common outcome aggregation scheme, and a rare outcome aggregation scheme. This methodology can be applied anywhere to aggregate geographic units and produce stable rates at a finer resolution.

Keywords

Aggregation; Census tract; Environmental health; Sub-county; Surveillance; Tracking

[*]Corresponding author at: ORISE Postdoctoral Fellow at the Environmental Public Health Tracking Section, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgia, United States, awerner@cdc.gov (A.K. Werner).

Authors' contributions: AKW contributed to the conception and design of the study and was responsible for the acquisition of data, geospatial work, statistical analyses, and interpretation of the results. AKW had primary responsibility for the manuscript. HS contributed to the conception and design of the study, interpretation of results, and editing of the manuscript. All authors have read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sste.2020.100339.

Availability of data and materials: Please contact corresponding author for data requests.

1. Introduction

Assessments have shown that timely, locally relevant information, including sub-county measures of health and associated factors, are important data needs for the public health community to have actionable data at a local level (Nagasako et al., 2018, DeSalvo et al., 2016, Castrucci et al., 2015). Despite what seems to be an explosion of data, technology, and computational power, public health data largely remain spatially unresolved, lagging, or all together unavailable. This underscores the importance of expanding the availability and accessibility of sub-county data not only for routine public health surveillance but also for decision-making and targeting interventions. The pressing need for sub-county data, including the need for community-level health indicators, has been echoed by many others (Nagasako et al., 2018, DeSalvo et al., 2016, Castrucci et al., 2015, Cutter et al., 1996, Drewnowski et al., 2013, Dwyer-Lindgren et al., 2017, Luck et al., 2006, Shah et al., 2014, Eisen and Eisen, 2007, Eisen and Eisen, 2008, Boothe et al., 2018).

Public health surveillance data resolved to the county level have been used to understand health burden, identify and target populations at-risk, and to guide and evaluate interventions. County health departments generally have administrative responsibility for the public residing in those counties. While useful, county-level data are limited (Courtemanche et al., 2015, Remington et al., 2015); in the way that state-level data can mask disparities across counties, county-level data can also mask considerable disparities at a local level (Drewnowski et al., 2013, Dwyer-Lindgren et al., 2017, CSTE 2017). Wide intra-county variations in socioeconomic indicators and risks [may] exist meaning that county-level data are useful, but local data are needed to measure and monitor what is happening (Cutter et al., 1996) and to more effectively target resources and interventions. Finer resolution data allow health departments to create community or neighborhood health profiles to better understand health issues affecting certain areas within a county, prioritize needs, and take action to advance policies and programs (Centers for Disease Control and Prevention 2013). These health departments need locally relevant, reliable indicators so they can have effective health promotion programs, provide better health services, and specific health planning and management (Rahman, 2017), and are not as limited in their ability to detect hot spots, identify determinants of health, and implement effective, targeted interventions (Boothe et al., 2018).

For environmental health, the need for local-level data is also driven by local-level heterogeneity in the environment. Counties often cover large areas and show environmental variability (Eisen and Eisen, 2008), which could mean variation in exposure and in habitat that are not uncovered without using finer geographic resolution data. While individual-level data would eliminate the problem, this introduces issues of confidentiality and stability. Often, this means that some form of aggregation, whether geographic or temporal, is applied before data are made available for public health practitioners, researchers, or the public. Aggregated data are commonly used in spatial epidemiologic studies because of data availability and/or other data limitations (e.g., lack of confounder and exposure data at the individual level) (Zhang et al., 2016, Beale et al., 2008), which allows for increased sample sizes and greater precision (Jia et al., 2004).

Established aggregations using administrative boundaries, such as census block groups (aggregations of census blocks) or census tracts (aggregations of block groups), can be used in sub-county analyses or surveillance. Typically, these spatial unit choices depend on data accessibility and availability, and census tracts are often the default unit for most studies where geography and health intersect (Lee et al., 2014). Census tracts are a reasonable solution for public health surveillance data as most addresses can be accurately geocoded to census tracts, population data are available for census tracts, and they are relatively stable over a ten-year period. However, numbers can still be too small for unsuppressed, stable displays, requiring further aggregation. Small numbers can create confidentiality issues when an area has a small population size (i.e., small denominator) and data reliability issues when an area has few cases (i.e., small numerator) (Colorado Department of Public Health and Environment 2017, VanEenwyk and Macdonald, 2012).

The increased availability of sub-county data would be useful to—and would better inform—a range of public and private organizations (Luck et al., 2006, Eisen and Eisen, 2008). In at least two states, legislators required the health department to release local-level health data to inform community action (Wu et al., 2018, New York State Department of Health 2017). With funding from the Robert Wood Johnson Foundation, CDC released census tract life expectancy measures and city-specific estimates of important behavioral risk factors (Centers for Disease Control and Prevention 2017, National Center for Health Statistics 2018). Lastly, the Institute of Medicine and the Public Health 3.0 initiative have called on the public health community and the CDC to make local-level data more widely available, accessible, and usable, linking environmental and human services data to health (DeSalvo et al., 2016, Office of the Assistant Secretary for Health 2016, Institute of Medicine 2012).

The Centers for Disease Control and Prevention's (CDC) Environmental Public Health Tracking Program (Tracking Program) is moving towards building a system of sub-county data to enhance the spatial resolution of data currently available in the National Environmental Public Health Tracking Network (Tracking Network). A previous project highlighted the need to address geographic aggregation and standardized geographies (Werner et al., 2018). One of the recommendations was to create generic aggregation schemes that would allow for tracking data over time (Werner et al., 2018).

While setting an aggregation scheme specifically for one outcome at one point in time would produce the best scheme for that outcome, it would result in numerous aggregation schemes over time. This means there would be different schemes for the many outcomes covered by the Tracking Network, making it difficult to manage and creating problems for comparisons over time and outcome. The goal of the Tracking Network is to provide integrated health, exposure, and environmental data (McGeehin et al., 2004); therefore, one or two standardized aggregation schemes that work well for multiple outcomes over time would allow for trend analysis and comparisons in this context. These geographies could be adopted by other entities for datasets not covered by the Tracking Network to expand the amount of data that could be comparable over time or to allow others in different localities to display their data at a finer resolution whilst protecting confidentiality and ensuring stability.

The aim of this project was to develop two aggregation schemes using census tracts as the foundation. One aggregation scheme would be for the Tracking Program's health outcomes that are considered more common and the other for the Tracking Program's health outcomes that are considered rarer. While there is no universal definition for rare health outcomes, they are generally defined as diseases affecting fewer than 1 in 2000 persons (Andermann, 2013) or having an average prevalence between 40 and 50 cases per 100,000 persons (Richter et al., 2015). Using these two aggregation schemes will allow for the dissemination of stable (i.e., reliable), unsuppressed data at the finest geography possible whilst enabling comparison of data across time, place, and outcome. The main objectives were to pilot a series of population thresholds for a number of states to determine which population thresholds work best and to recommend parameters to guide these aggregation schemes.

2. Methods

2.1. Background

The Tracking Program worked on creating standardized geographies using census tracts as the foundation, ensuring that the maximum number of sub-county geographies could be displayed with minimal suppression and instability. While standardized geographies will ultimately be created for all states, a subset of Tracking Program recipients were involved with this project as recipients elect to participate in the workgroups. Health outcome data used in this project represent what are considered more common and rarer outcomes in the context of available Tracking Program data. Some recipients did not have data available for both of the selected outcomes. These factors guided the states and datasets that were piloted.

2.2. Data

Population data were acquired from the U.S. Census Bureau 2010 decennial census for population by age and sex categorized by census tract (U.S. Census Bureau 2017). County-level asthma emergency department (ED) data, which recipient states routinely submit to the Tracking Program, were selected for six recipients (Florida, Maine, Minnesota, New Hampshire, New York, including New York City, and Wisconsin). This outcome represented one of the Tracking Program's more common outcomes. County-level lung and bronchus cancer data, which come from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program and CDC's National Program of Cancer Registries, were obtained for Colorado, Florida, Maine, Missouri, New Hampshire, New York, including New York City, and Wisconsin. These data represented one of the Tracking Program's rarer outcomes. All health outcome datasets used 2010 data.

2.3. Calculating expected census tract-level counts

At present, privacy concerns prevent many Tracking recipients from submitting census tract-level data to the Tracking Program without amending data use agreements. To overcome this and to complete this pilot, annual expected case counts were calculated for 2010. SAS 9.3 was used for creating expected count datasets. Census tract-level population data were aggregated to their respective counties within each state to calculate population totals for each county by 5-year age group and gender. Case count data for each health outcome were merged with the county-level population data, and county-level age- and sex-specific

rates were calculated. County-level rate data were applied to census tract-level population data. Expected counts were then calculated for each census tract (county rate * census tract population) by age and gender. Table 1 provides an example of the expected case count data for one census tract in a given county (with modified values), which shows how the age and sex-specific county rates were applied to the census tract-level population data to obtain the expected case counts.

Final datasets were exported for each health outcome containing census tract geographic IDs, total population, population counts for those in the 0–4 year and 65+ year age groups, and total expected case counts. Any census tracts with a population of zero were removed to avoid aggregating these areas, which could artificially increase the size of an aggregated area and skew the statistics.

2.4. Geographic aggregation

Census tract shapefiles were downloaded for each state from the U.S. Census Bureau (U.S. Census Bureau 2017), and the files were imported into ArcGIS. The census tract-level expected count tables were joined to the census tract shapefiles on matching geographic IDs, resulting in one shapefile with 2010 tract boundaries, populations, and expected counts for the selected health outcomes. This shapefile was used as the input for the Geographic Aggregation Tool (GAT), an R or SAS program, which facilitates the creation of compact geographic units that meet criteria specified by the user (Talbot and LaSelva, 2010). The R version was used for this work.

Several population thresholds for aggregating census tracts were tested based on either total population or based on a threshold sum of a sub-population (0–4 and 65+ year olds) with preference given to within county aggregations. The sub-population was chosen as an alternative threshold compared to the total population because it was thought that this would be a marker for the overall population threshold (i.e., once the sub-population of 0–4 and 65+ year olds combined reached a certain size together, then it should be stable for the total population as these age groups tend to be smaller).

Total population thresholds tested included 5000, 10,000, 15,000, 20,000, and 25,000 persons. For sub-population aggregation, 2010 Census data were examined to determine the population distribution for 0–4 and 65+ year olds (Howden and Meyer, 2011). The population for 0–4 year olds was 20,201,362 persons and the population for 65+ year olds was 40,267,984 persons; therefore, the population thresholds in the GAT were set so that the 65+ year old age group was twice as much as the 0–4 year old age group. The sub-population thresholds were as follows: 1000 persons (333 persons 0–4 years old and 667 persons 65+ years old); 2500 persons (833 persons 0–4 years old and 1667 persons 65+ years old); and 500 0 persons (1667 persons 0–4 years old and 3333 persons 65+ years old).

A new shapefile representing a new aggregation level was produced by the GAT for each health outcome, state, and population threshold, resulting in eight new shapefiles for each state for each health outcome. Excel files were imported into SAS from the GAT-created shapefiles.

Readers can dynamically view the geographies via the Tracking Network's Data Explorer (https://ephtracking.cdc.gov/DataExplorer/#/). Census data can be viewed and evaluated to compare the demographic and socioeconomic composition of the proposed geographies compared to census tract and county.

2.5. Statistical calculations

SAS 9.3 was used for statistical calculations. Prevalence, confidence intervals (CI), percent suppressed, and percent unstable were calculated for each geographic unit in census tracts and in each new aggregation level. Relative standard error (RSE) was used to measure the stability or reliability of the data and was calculated as follows: (standard error/prevalence rate) * 100 where standard error = prevalence rate/√(expected cases). So, RSE = 1/√(expected cases)*100 (New York State Department of Health 1999). If there were no cases, then the RSE was set to missing. Each geographic unit was flagged for suppression if the number of expected cases was more than zero but less than 6 (following the Tracking Network's current rule). A RSE of <30% was considered acceptable and stable. Likewise, a threshold of <30% was considered acceptable for the number of suppressed geographic units.

The expected case count distribution was examined for census tract and for each new aggregation level for each health outcome for each state. A series of scatterplots were created to examine the number of expected cases relative to the RSE for each aggregation level. Statistics were calculated to determine how the median population changed across aggregation levels. The percent of suppressed and unstable geographies were reviewed for each of the population thresholds. The results were examined to look for natural cut-points where suppression and instability were minimized (i.e., <30%) and the number of geographic units was maximized, balancing the need for data reliability and confidentiality. This means that the optimal population threshold was where both suppression and instability were generally less than 30%, selecting as low of a population threshold as possible, which resulted in more geographic units.

2.6. Refining median case count ranges

After the new aggregation levels were reviewed, an optimal population threshold was selected for the rarer outcome aggregation scheme and for the more common outcome aggregation scheme. The median expected case counts for each state were examined for each health outcome to recommend an aggregation scheme based on a median case count range. For asthma, expected cases were subtracted from the states with the lowest median case count to determine the lower limit of the median case count range for the common aggregation scheme. Expected cases were also subtracted from states with higher median case counts to determine the upper limit of the median case count range for the common aggregation scheme. The upper limit would also serve as the cut-point for which census tract-level data should be stable. For lung and bronchus cancer, cases (or a fraction thereof) were added to the state with the lowest median case count to decide on the lower limit of the median case count range that would be acceptable for the rare aggregation scheme.

3. Results

Asthma ED served as a suitable Tracking Program outcome to represent the more common aggregation scheme. Fig. 1 shows the RSE as a function of asthma ED expected cases across each population threshold for Florida. Each point on the graph represents one aggregated sub-county geographic area. A reference line indicates the 30% RSE threshold, with the goal of identifying the population threshold where the majority of the points were below the reference line. The figures for the remaining pilot states are shown in Supplemental Figure S1, which showed that generally, the RSE was acceptable (i.e., below 30) for the population threshold of 5000 persons across all of the pilot states.

Table 2 provides more details, showing the descriptive statistics for the expected case counts for each aggregation level, including the original census tract-level data, as well as the total number of geographic units and the percentage of geographic units that were suppressed or unstable. Census tract-level median asthma ED expected case counts ranged from 12.8 (NH) to 25.1 (NY) cases. Table 2 shows how the percent unstable column decreases with increasing case counts across the different aggregation levels.

Without any aggregation, the percentage of unstable geographic units ranged from 10.2% (NY) to 38.0% (MN). The lowest level of aggregation (total population of 5000 persons) reduced instability to a range of 0% to 6.2% unstable, showing a marked decrease in the number of geographic units that would be flagged as unstable. Likewise, the percentage of census tract-level suppressed geographic units ranged from 1.4% to 11.3%, which was reduced to 0% to 5.6% suppressed using an aggregation level of 5000 persons. Asthma ED prevalence remained fairly similar across aggregation levels and increasing the aggregation level resulted in tighter confidence intervals due to increasing stability, which was expected.

The lower and upper limits of the median expected case count range were tested to determine the optimal range for which the common outcome aggregation scheme should be used. For example, New Hampshire had the lowest median case count range (12.8 cases), so expected cases were subtracted from this value to test where the aggregation level of 5000 persons was no longer stable. Additional testing for the lower limit of a suitable median case count range for the more common aggregation scheme showed somewhere between 6.8 and 7.8 cases as acceptable, with a lower end of approximately 7.3 cases recommended for this scheme.

Testing for the upper limit of a suitable median case count range showed approximately 16.9 cases as suitable based on the percentage of geographic units flagged as suppressed and unstable. This was determined by subtracting expected cases from those states that had higher median case counts at the census tract level (e.g., Florida) to determine where the aggregation level of 5000 persons started to lose stability. Therefore, any health outcomes with a median case count of 17.0 cases or higher at the census tract level should be acceptable in terms of stability and suppression.

Lung and bronchus cancer served as a suitable Tracking Program outcome to test the rare outcome aggregation scheme, with expected census tract-level median case counts ranging from 1.6 (CO) to 3.6 (NH) cases. Fig. 2 shows the RSE as a function of lung and bronchus

cancer expected cases across each population threshold for New Hampshire. The figures for the remaining pilot states are shown in Supplemental Figure S2. Generally, the RSE was acceptable (i.e., below 30) for the population threshold of 20,000 persons.

More details are shown in Table 3, which include lung and bronchus cancer case count data, the number of geographic units for each threshold, and the corresponding percentage of geographic units that were suppressed or unstable. Census tract-level data showed the percentage of geographic units (i.e., census tracts) that were flagged as unstable ranged from 98.3% to 100%, indicating that census tract-level data would not be suitable to display due to small numbers and related data reliability issues. Census tract-level suppression ranged from 84% to 98.7% of geographic units suppressed. As the aggregation level increases, the percent unstable column decreases.

As expected, the percentage of unstable geographic units decreased with increasing aggregation. Using the total population aggregation of 20,000 persons, the percentage of unstable geographic units were reduced to 0.0% to 39.4% unstable and the percentage of suppressed geographic units was reduced to 0.0% to 5.0% suppressed. As with the more common outcome, lung and bronchus cancer prevalence was similar across aggregation thresholds, as expected, and increasing the aggregation levels resulted in tighter confidence intervals. The highest percentage of unstable geographies was attributed to Colorado, which had a median census tract-level case count of 1.6 cases, whereas the other states ranged from 2.4 to 3.6 cases.

The lower limit of the common outcome aggregation scheme (i.e., 7.3 median expected case count) dictated the upper limit for the rare outcome aggregation scheme. Therefore, a median case count of 7.2 was a suitable upper limit for the rare outcome aggregation scheme. Additional testing for the lower limit of a suitable median case count range for the rare outcome aggregation scheme showed between 1.8 and 2.0 cases as acceptable, with a lower end of approximately 1.9 cases recommended for this scheme. This was determined by adding fractions of expected cases to the state with the lowest median expected case count (i.e., Colorado with a median case count of 1.6 cases) to determine where the 20,000 person aggregation started to become more stable. It is suggested that any health outcomes with a median case count of less than 1.9 use temporal aggregation in conjunction with geographic aggregation to achieve stability.

Table 4 summarizes the recommended aggregation scheme, the corresponding median case count ranges, and the population thresholds when looking at a single year of data.

4. Discussion

The aim of this study was to develop two aggregation schemes, including one for rarer health outcomes and one for more common health outcomes, by piloting geographic aggregation work with several Tracking recipients and examining data stability and confidentiality issues. Ultimately, a range of median case counts was suggested for each aggregation scheme with the goal of having data that are comparable across time, health outcomes, and location. Any health outcomes with a census tract-level median case count

of 17.0 cases or higher should be appropriate for these data to be displayed without aggregation. The suggested common outcome aggregation scheme, using a total population of 5000 persons, had a census tract-level median case count range of 7.3 to 16.9 cases. The suggested rare outcome aggregation scheme, using a total population of 20,000 persons, had a census tract-level median case count range of 1.9 to 7.2 cases.

Generally, most studies that have examined the link between local place effects and health impacts have used administrative units (i.e., census areas) for the geographies, typically due to the availability of underlying population data (Haynes et al., 2007). This was also done in this study, using pre-defined administrative units (i.e., census tracts) as the base geography for further geographic aggregation. In determining which aggregation levels worked best for the two schemes, it was important to balance data stability and confidentiality issues whilst keeping the geographic areas at the highest possible resolution.

The primary statistical issue with small areas is that of data stability and reliability because of too few cases (National Association of Health Data Organizations 2004, Brownson et al., 2010). When calculating rates, a larger numerator will allow for a rate to better estimate the true/underlying rate of the population (Buescher, 2008). However, obtaining sufficient numerators becomes problematic when using small areas, especially for rare outcomes. This was seen with the differences in the total population required for each aggregation scheme (i.e., 5000 persons versus 20,000 persons). There may also be times when the numbers are too small and will require too much temporal and/or spatial aggregation to be useful. With smaller numbers, it is desirable to have accompanying confidence intervals to understand the uncertainty around point estimates (National Association of Health Data Organizations 2004). This will be important for the Tracking Program to incorporate into displays when shifting to sub-county data on the Tracking Network so users can better understand associated uncertainties.

With this work, it is also important to recognize the modifiable areal unit problem (MAUP). Grouping the data at different spatial resolutions (e.g., census tracts, aggregations of census tracts, county) or grouping the data in different ways at the same scale can lead to variation in the results and subsequent interpretation of those results (Beale et al., 2008). While the MAUP remains unsolved (Zhang et al., 2016), generally, analysis is recommended for the smallest area units with available data, and aggregation to larger units should be avoided unless there is appropriate rationale (Pfeiffer et al., 2008). For this study, the smallest area units (i.e., census tracts) were used, and aggregating to larger units was necessary to calculate stable estimates and protect confidentiality, with the level of aggregation depending on census tract-level median case counts. Without some form of aggregation, little to no data would be displayed on the Tracking Network for most health outcomes at a sub-county level, particularly for rarer health outcomes.

The creation of these sub-county geographies used geographic aggregation without modelling, smoothing, or other estimation methods, which was chosen based on census tract-level data availability, recipient input, and to facilitate user understanding and ease of interpretation of the data when accessing the Tracking Network. Methodologies are available for small area estimation, with the two major approaches being Bayesian

modelling and spatial microsimulation (Koh et al., 2018). Spatial microsimulation creates synthetic population datasets for small areas where existing data are unavailable (Rahman et al., 2010). Spatial interpolation and smoothing (e.g., Bayesian modelling) use nearby observations to fill in and/or improve estimates, which improves rates in areas with few observations (Auchincloss et al., 2012). In one study, hierarchical Bayesian models were used to estimate the prevalence of current smoking; however, the authors noted that relatively low precision was obtained from these models and suggested aggregating census tracts into larger areas (as in this study) or increasing temporal aggregation (Song et al., 2016). For this work, we chose to use one of several spatial epidemiology methods (Auchincloss et al., 2012), which was the use of aggregation.

Preliminary geographic aggregation work for the rare outcome aggregation scheme explored health outcomes rarer than lung and bronchus cancer; however, these were not suitable for use without temporal aggregation. Liver cancer was initially selected to test the rare outcome scheme, which resulted in nearly all aggregation levels to show 100% of geographies as suppressed. Non-Hodgkin's lymphoma was then tested, but this still showed the majority of aggregation levels as suppressed. Therefore, the work presented here used lung and bronchus cancer to represent rarer outcomes. In 2014, some states had lung and bronchus cancer incidence rates of 50 cases or fewer per 100,000 persons (Centers for Disease Control and Prevention 2017).

This illustrates the fact that there are certain health outcomes (i.e., those with a median case count less than 1.9 cases) where the proposed aggregation schemes will not work for annual data and temporal aggregation will be required. However, temporal aggregation would not allow for examination of time-trend differences (Jia et al., 2004). Additionally, any area differences that are noted where geographic aggregation is used would require spatial delineation, where possible, to explore these differences further (Jia et al., 2004).

This pilot project used readily available data to test and propose geographic aggregation schemes for sub-county work. While this work relied on calculating expected counts for census tracts, these aggregation schemes should be tested further using census tract-level data from all Tracking recipients, where data are available. Recipient access to these data may require changes to data use agreements. Future refinement of this work includes testing nested aggregation schemes (i.e., aggregating the more common outcome aggregation scheme within the rare outcome aggregation scheme) to have a hierarchical structure of geographies and removing more rural areas where county boundaries may be crossed. Different aggregation methods will be explored to account for certain geographic features (e.g., waterbodies) or population density. Rural and urban differences could also be investigated, determining if different aggregation methods should be used so as to avoid grouping these areas together. Additional work could include accounting for other factors, such as socioeconomic status or race/ethnicity, when aggregating geographies based on the suggested aggregation schemes.

This work has several limitations. In calculating expected census tract-level case counts, it was assumed that the prevalence was the same across all census tracts in a county because of applying county-level rates to census tracts. This likely results in variation

from actual census tract-level case counts, but the methodology is still applicable and will be tested using census tract-level datasets when they can be obtained. Expected counts were calculated using 2010 decennial census data. Certain states, or areas within states, that experienced significant population growth or decline since the 2010 decennial census may have expected counts that are not necessarily reflective of the present; however, there are limitations, such as larger margins of error, associated with using other data sources (e.g., American Community Survey estimates). Not all states included in the pilot had data available for both datasets, so we were unable to test all of the states for both aggregation schemes. Additionally, the outcomes that are defined as rare or common in this paper are rare or common in the context of Tracking data and may not be defined that way elsewhere. It should also be noted that densely populated cities could influence the median case counts of a state. For example, the median case count for asthma ED visits for New York State was heavily influenced by the case counts coming from New York City.

Preliminary examination of publicly available finer geographic resolution data, such as the sub-county data that will be displayed on the Tracking Network in the future, can help to develop novel hypotheses to explore further, particularly with more detailed datasets (Boscoe et al., 2015). The methods presented in this pilot project are applicable more broadly and demonstrate how geographic aggregation can be used for surveillance purposes beyond the Tracking Program in the United States to achieve stable estimates whilst considering confidentiality issues. There are limitations associated with these methods and the display of data; however, the goal is to disseminate data at the sub-county level to lead to more public health actions at the local level. Future work will help to refine these methods, the aggregation schemes, and the usefulness of data being presented to users. Ultimately, disseminating sub-county data (or finer resolution data in other countries) can help address local environmental health decision-making, identify local variation, advance understanding of environmental health processes and impacts, improve surveillance, and target interventions (Castrucci et al., 2015).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank the Environmental Public Health Tracking Program’s Geospatial Standards and Network Development Workgroup (Kevin Berg (CO) and Craig Kassinger (CDC) co-leads) for all of their feedback throughout this process and would like to thank those who reviewed the geographies: Kevin Berg (Colorado), Chris DuClos (Florida), Chris Paulu (Maine), Jessie Shmool (Minnesota), Jeff Patridge (Missouri), Katie Bush and Jessie Sagona (New Hampshire), Doug Done, Neil Muscatiello, and Arjita Rai (New York State), and Jenny Camponeschi and Paul Creswell (Wisconsin). The authors would also like to thank Beverly Levine (Wake Forest University), Douglas Levine (University of North Carolina), and Mikyong Shin (Tracking Program) for their review of the draft manuscript.

This work was supported in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

Disclaimer

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.

Abbreviations:

CDC	Centers for Disease Control and Prevention
CI	confidence interval
ED	emergency department
GAT	Geographic Aggregation Tool
MAUP	modifiable areal unit problem
RSE	relative standard error
Tracking Network	National Environmental Public Health Tracking Network
Tracking Program	National Environmental Public Health Tracking Program

References

Andermann A, 2013. Evidence for Health: From Patient Choice to Global Policy. Cambridge University Press, New York, New York.

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

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

Boothe VL, Fierro LA, Laurent A, Shih M, 2018. Sub-county life expectancy: a tool to improve community health and advance health equity. Prev. Chronic Dis 15, E11. [PubMed: 29369759]

Boscoe FP, Talbot TO, Kulldorff M, 2015. Public domain small-area cancer incidence data for New York State, 2005-2009. University at Albany Scholars Archive, Albany, NY Available at: http://scholarsarchive.library.albany.edu/cgi/viewcontent.cgi?article=1004&context=epi_fac_scholar.

Brownson RC, Baker EA, Leet TL, Gillespie KN, True WR, 2010. Evidence-Based Public Health, 2nd edn Oxford University Press, Inc., Oxford.

Buescher PA, 2008. Problems with rates based on small numbers. North Carolina Public Health State Center for Health Statistics, Raleigh, NC Available at: http://www.schs.state.nc.us/schs/pdf/primer12_2.pdf.

Castrucci BC, Rhoades EK, Leider JP, Hearne S, 2015. What gets measured gets done: an assessment of local data uses and needs in large urban health departments. J. Public Health Manag. Practice 21, S38–S48.

Centers for Disease Control and Prevention. Creating a health profile of your neighborhood: a how-to guide. [https://www.cdc.gov/healthyplaces/toolkit/sources_of_health_data.pdf] (2013). Accessed January 6 2020.

Centers for Disease Control and Prevention. 500 Cities: local data for better health. [https://www.cdc.gov/500cities/index.htm] (2017). Accessed January 14 2019.

Centers for Disease Control and Prevention. Lung cancer rates by state. [https://www.cdc.gov/cancer/lung/statistics/state.htm] (2017). Accessed October 24 2017.

Colorado Department of Public Health and Environment. Guidelines for working with small numbers. [http://www.cohid.dphe.state.co.us/smnumguidelines.html]. Accessed 7 March 2017.

Courtemanche C, Soneji S, Tchernis R, 2015. Modeling area-level health rankings. Health Serv. Res 50, 1413–1431, [PubMed: 26256684]

Cutter SL, Holm D, Clark L, 1996. The role of geographic scale in monitoring environmental justice. Risk Anal. 16, 517–526.

CSTE. Guide for sub-county assessment of life expectancy (SCALE). 2017. Available at: https://cdn.ymaws.com/www.cste.org/resource/resmgr/pdfs/pdfs2/SCALE_Report_v1.pdf.

DeSalvo KB, O'Carroll PW, Koo D, Auerbach JM, Monroe JA, 2016. Public health 3.0: time for an upgrade. Am. J. Public Health 106, 621–622. [PubMed: 26959263]

Drewnowski A, Rehm CD, Arterburn D, 2013. The geographic distribution of obesity by census tract among 59 767 insured adults in king county, WA. Int. J. Obes 38, 833.

Dwyer-Lindgren L, Stubbs RW, Bertozzi-Villa A, Morozoff C, Callender C, Fine-gold SB, Shirude S, Flaxman AD, Laurent A, Kern E, et al. , 2017. Variation in life expectancy and mortality by cause among neighbourhoods in king county, WA, USA, 1990-2014: a census tract-level analysis for the global burden of disease study 2015. Lancet Public Health 2, e400–e410. [PubMed: 29253411]

Eisen L, Eisen RJ, 2007. Need for improved methods to collect and present spatial epidemiologic data for vectorborne diseases. Emerg. Infect. Dis 13, 1816–1820, [PubMed: 18258029]

Eisen RJ, Eisen L, 2008. Spatial modeling of human risk of exposure to vector-borne pathogens based on epidemiological versus arthropod vector data. J. Med. Entomol 45, 181–192. [PubMed: 18402133]

Haynes R, Daras K, Reading R, Jones A, 2007. Modifiable neighbourhood units, zone design and residents' perceptions. Health Place 13, 812–825. [PubMed: 17369075]

Howden LM, Meyer JA, 2011. Age and sex composition: 2010. U.S. Department of Commerce, U.S. Census Bureau, Washington, D.C. Available at: https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.

Institute of Medicine. For the public's health: investing in a healther future. [http://www.nationalacademies.org/hmd/Reports/2012/For-the-Publics-Health-Investing-in-a-Healthier-Future.aspx] (2012). Accessed.

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

Koh K, Grady SC, Darden JT, Vojnovic I, 2018. Adult obesity prevalence at the county level in the united states, 2000–2010: downscaling public health survey data using a spatial microsimulation approach. Spat. Spatio-temporal Epidemiol 26, 153–164.

Lee J, Alnasrallah M, Wong D, Beaird H, Logue E, 2014. Impacts of scale on geographic analysis of health data: an example of obesity prevalence. ISPRS Int. J. Geoinf 3, 1198–1210,

Luck J, Chang C, Brown ER, Lumpkin J, 2006. Using local health information to promote public health. Health Aff 25, 979–991,

McGeehin MA, Quakers JR, Niskar AS, 2004. National environmental public health tracking program: bridging the information gap. Environ. Health Perspect 112, 1409–1413. [PubMed: 15471734]

Nagasako E, Waterman B, Reidhead M, Lian M, Gehlert S, 2018. Measuring sub-county differences in population health using hospital and census-derived data sets: the Missouri zip health rankings project. J. Public Health Manag. Practice 24, 340–349.

National Center for Health Statistics. U.S. small-area life expectancy estimates project (USALEEP): life expectancy estimates files, 2010-2015. [https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html] (2018). Accessed.

National Association of Health Data Organizations. Statistical approaches for small numbers: addressing reliability and disclosure risk. NAhDO-CDC cooperative agreement project: 2004. Available at:https://www.nahdo.org/sites/nahdo.org/files/Data_Release_Guidelines.pdf.

New York State Department of Health. Rates based on small numbers - statistics teaching tools. [https://www.health.ny.gov/diseases/chronic/ratesmall.htm] (1999). Accessed March 7 2017.

New York State Department of Health. Cancer incidence by census tract. [https://www.health.ny.gov/statistics/cancer/registry/tract/index.htm] (2017). Accessed 10 January 2019.

Office of the Assistant Secretary for Health. Public health 3.0: a call to action to create a 21st century public health infrastructure. [https://www.healthypeople.gov/sites/default/files/Public-Health-3.0-White-Paper.pdf?_ga=2.199279278.1689288573.1530116058-486461828.1530116058] (2016). Accessed.

Pfeiffer D, Robinson T, Stevenson M, Stevens K, Rogers D, Clements A, 2008. Spatial Analysis in Epidemiology. Oxford University Press, New York, New York.

Rahman A, Harding A, Tanton R, Liu S, 2010. Methodological issues in spatial microsimulation modelling for small area estimation. Int. J. Microsimulat 3, 3–22.

Rahman A, 2017. Estimating small area health-related characteristics of popula-tions: a methodological review. Geospat. Health 12, 495. [PubMed: 28555467]

Remington PL, Catlin BB, Gennuso KP, 2015. The county health rankings: rationale and methods. Popul. Health Metr 13, 11. [PubMed: 25931988]

Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, Hughes DA, 2015. Rare disease terminology and definitions—A systematic global review: report of the ISPOR rare disease special interest group. Value in Health 18, 906–914. [PubMed: 26409619]

Shah SN, Russo ET, Earl TR, Kuo T, 2014. Measuring and monitoring progress toward health equity: local challenges for public health. Prev. Chronic Dis 11, E159. [PubMed: 25232746]

Song L, Mercer L, Laurent A, Solet D, 2016. Using small-area estimation to calculate the prevalence of smoking by subcounty geographic areas in king county, Washington, behavioral risk factor surveillance system, 2009-2013. Prev. Chronic Dis 5, E59.

Talbot TO, LaSelva GD, 2010. Geographic Aggregation Tool, Version 1.31, New York State Health Department, Troy, NY.

U.S. Census Bureau. American Factfinder. [https://factfinder.census.gov/] (2017). Accessed June 2 2017.

U.S. Census Bureau. Cartographic boundary shapefiles - census tracts. [https://www.census.gov/geo/maps-data/data/cbf/cbf_tracts.html] (2017). Accessed July 11 2017.

VanEenwyk J, Macdonald SC, 2012. Guidelines for working with small numbers. Washington State Department of Health, Tumwater, WA Available at: http://www.doh.wa.gov/Portals/1/Documents/5500/SmallNumbers.pdf.

Werner AK, Strosnider H, Kassinger C, Shin M, 2018. Lessons learned from the environmental public health tracking sub-county data pilot project. J. Public Health Manag. Practice 24, E20–E27.

Wu XC, Maniscalco L, Zhang L, Yi Y, Lefante C, Ricks L, Straif-Bourgeois S, Hsieh MC Cancer incidence in Louisiana by census tract, 2006–2014. https://sph.lsuhsc.edu/wp-content/uploads/2019/01/01_Cancer-Incidence-in-Louisiana-by-Census-Tract-2006-2014.pdf (2018). Accessed.

Zhang Z, Manjourides J, Cohen T, Hu Y, Jiang Q, 2016. Spatial measurement errors in the field of spatial epidemiology. Int. J. Health Geogr 15, 21. [PubMed: 27368370]

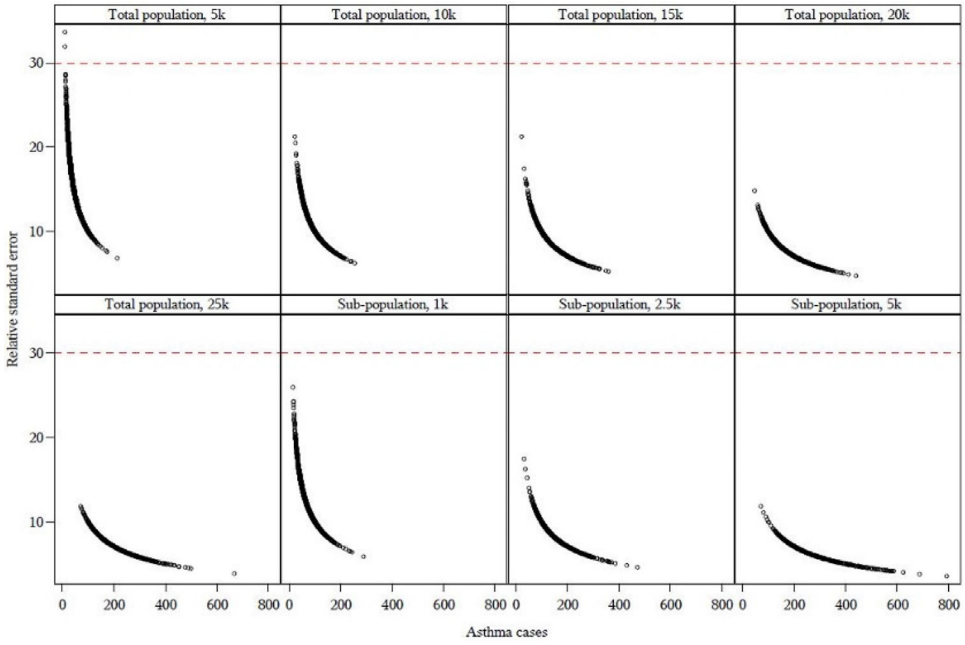

Fig. 1.

Relative standard error (RSE) as a function of expected census tract-level asthma emergency department cases for Florida. Each point on the graph represents one aggregated sub-county geographic area. The reference line is where RSE=30, with stable RSEs displayed below the line. See Figure S1 for graphs of the remaining states.

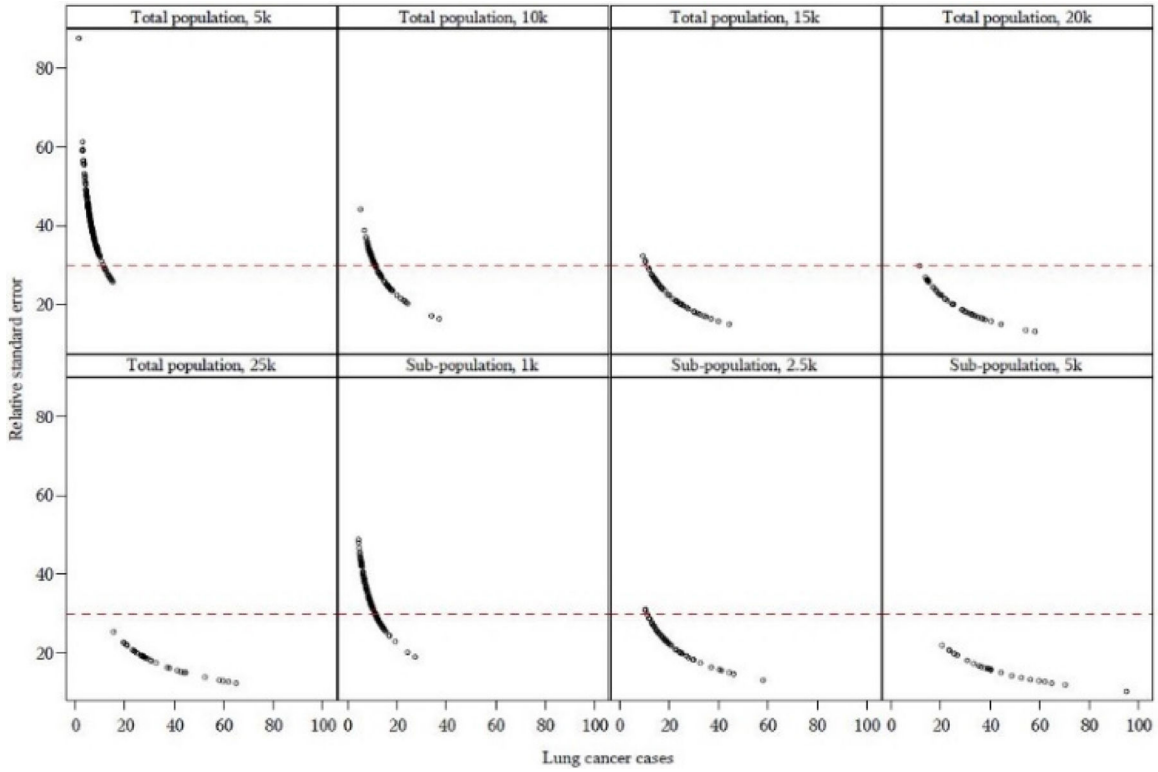

Fig. 2.

Relative standard error (RSE) as a function of expected census tract-level lung and bronchus cancer cases for New Hampshire. Each point on the graph represents one aggregated sub-county geographic area. The reference line is where RSE=30, with stable RSEs displayed below the line. See Figure S2 for graphs of the remaining states.

Table 1

An example of census tract data showing how the county rate is applied to each age group and gender at the census tract level to obtain the expected case counts.

Geo ID	State	County	Gender	Age group (years)	County rate	Population	Census tract expected case count*
12086000107	FL	Miami-Dade	F	0–4	0.046	44	2.02
12086000107	FL	Miami-Dade	F	5–9	0.009	37	0.33
12086000107	FL	Miami-Dade	F	10–14	0.012	47	0.56
12086000107	FL	Miami-Dade	F	15–19	0.002	36	0.07
12086000107	FL	Miami-Dade	F	20–24	0.015	75	1.13
12086000107	FL	Miami-Dade	F	25–29	0.008	129	1.03
12086000107	FL	Miami-Dade	F	30–34	0.011	86	0.95
12086000107	FL	Miami-Dade	F	35–39	0.002	85	0.17
12086000107	FL	Miami-Dade	F	40–44	0.010	103	1.03
12086000107	FL	Miami-Dade	F	45–49	0.004	101	0.40
12086000107	FL	Miami-Dade	F	50–54	0.009	91	0.82
12086000107	FL	Miami-Dade	F	55–59	0.016	75	1.20
12086000107	FL	Miami-Dade	F	60–64	0.021	70	1.47
12086000107	FL	Miami-Dade	F	65–69	0.036	59	2.12
12086000107	FL	Miami-Dade	F	70–74	0.014	47	0.66
12086000107	FL	Miami-Dade	F	75–79	0.008	39	0.31
12086000107	FL	Miami-Dade	F	80–84	0.012	21	0.25
12086000107	FL	Miami-Dade	F	85+	0.015	28	0.42

**Note:* The census tract expected case count was obtained by multiplying the county rate by the census tract population. County rate and expected case count values have been changed for purposes of presenting this table.

Table 2

Descriptive statistics for asthma emergency department data by aggregation level and state showing expected case counts, number of geographic units, suppression, and stability.

Threshold	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable
Census tract		0.003	22	212	4181	5	17		2	13	39	292	11	35
Total, 5k		**9**	**43**	**212**	**2222**	**0**	**0.09**		**3**	**22**	**65**	**161**	**6**	**6**
Total, 10k		22	82	254	1185	0	0		9	42	159	83	0	1
Total, 15k		22	122	359	798	0	0		13	68	161	50	0	0
Total, 20k	*Florida*	46	158	441	611	0	0	*New Hampshire*	12	75	258	42	0	0
Total, 25k		71	200	668	493	0	0		39	97	310	32	0	0
Sub, 1k		15	62	288	1560	0	0		4	31	115	114	4	6
Sub, 2.5k		33	136	472	706	0	0		12	68	258	49	0	0
Sub, 5k		71	264	791	365	0	0		39	127	417	25	0	0
Census tract		2	21	57	351	1	10		0.003	25	593	4870	2	10
Total, 5k		**22**	**44**	**152**	**159**	**0**	**0**		**7**	**53**	**593**	**2357**	**0**	**0.3**
Total, 10k		47	88	233	80	0	0		13	102	790	1237	0	0
Total, 15k		69	151	377	51	0	0		26	150	948	862	0	0
Total, 20k	*Maine*	110	183	429	39	0	0	*New York*	55	206	1256	642	0	0
Total, 25k		119	216	757	31	0	0		55	253	1598	505	0	0
Sub, 1k		22	59	214	119	0	0		7	69	675	1752	0.	0.3
Sub, 2.5k		80	163	377	44	0	0		16	156	1676	764	0	0
Sub, 5k		140	322	762	23	0	0		55	342	2413	373	0	0
Census tract		0.5	14	77	1334	9	38		0	15	96	1392	7	34
Total, 5k		**2**	**29**	**101**	**630**	**0.5**	**2**		**3**	**26**	**151**	**683**	**0.3**	**4**
Total, 10k		4	56	202	332	0.3	0.9		15	53	246	355	0	0
Total, 15k		4	85	229	227	0.4	0.4		27	84	323	238	0	0
Total, 20k	*Minnesota*	17	111	318	171	0	0	*Wisconsin*	27	104	458	181	0	0
Total, 25k		23	143	338	134	0	0		47	141	676	145	0	0
Sub, 1k		2	38	167	470	0.6	2		6	33	338	520	0	1
Sub, 2.5k		4	89	319	199	0.5	0.5		27	81	466	227	0	0
Sub, 5k		17	145	571	117	0	0		57	157	1326	115	0	0

Note: All case counts shown are expected census tract-level case counts. Any numbers greater than 1 have been rounded to the nearest whole number and are referred to in the text by the complete value. Total population 5000 persons is the recommended aggregation scheme for common outcomes. Sub-population aggregation thresholds were as follows: 1000 persons (333 persons 0–4 years old and 667 persons 65+ years old); 2500 persons (833 persons 0–4 years old and 1667 persons 65+ years old); and 5000 persons (1667 persons 0–4 years old and 3333 persons 65+ years old).

Spat Spatiotemporal Epidemiol. Author manuscript; available in PMC 2022 February 14.

Table 3

Descriptive statistics for lung and bronchus cancer by aggregation level and state showing expected case counts, number of geographic units, suppression, and stability.

Threshold	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable
Census tract		0	2	9	1242	99	100		0.5	4	10	292	95	100
Total, 5k		0	3	17	606	89	99		1	6	15	161	51	91
Total, 10k		0.1	6	25	316	47	89		5	11	37	83	1	51
Total, 15k		3	9	30	220	19	69		10	20	44	50	0	6
Total, 20k	***Colorado***	**4**	**13**	**42**	**160**	**5**	**39**	***New Hampshire***	**11**	**22**	**58**	**42**	**0**	**0**
Total, 25k		4	15	42	134	3	26		15	28	65	32	0	0
Sub, 1k		0.4	5	17	407	68	96		4	9	27	114	17	72
Sub, 2.5k		3	12	45	170	2	44		10	18	58	49	0	4
Sub, 5k		7	24	51	87	0	2		21	39	95	25	0	0
Census tract		0	3	66	4181	84	98		0	2	23	4870	95	100
Total, 5k		0.009	6	72	2222	46	84		0.03	5	23	2357	61	95
Total, 10k		1	12	76	1185	11	47		0.9	10	44	1237	16	61
Total, 15k		4	18	101	798	2	18		2	14	55	862	4	33
Total, 20k	***Florida***	**4**	**23**	**145**	**611**	**1**	**9**	***New York***	**4**	**19**	**78**	**642**	**0.6**	**16**
Total, 25k		8	30	141	493	0	2		7	24	81	505	0	6
Sub, 1k		2	8	82	1560	25	68		3	7	34	1752	38	85
Sub, 2.5k		7	19	165	706	0	14		7	16	75	764	0	24
Sub, 5k		14	36	236	365	0	0		13	32	101	373	0	0
Census tract		0.4	3	10	351	92	100		0.003	3	9	1392	97	100
Total, 5k		3	7	17	159	26	84		0.1	5	26	683	59	93
Total, 10k		7	15	32	80	0	21		3	10	40	355	10	58
Total, 15k		10	23	55	51	0	2		2	16	53	238	2	20
Total, 20k	***Maine***	**15**	**29**	**65**	**39**	**0**	**0**	***Wisconsin***	**5**	**20**	**55**	**181**	**0.6**	**4**
Total, 25k		17	37	94	31	0	0		10	25	60	145	0	2
Sub, 1k		4	9	26	119	8	62		2	7	40	520	38	85
Sub, 2.5k		14	27	55	44	0	0		7	16	44	227	0	18
Sub, 5k		24	55	100	23	0	0		16	33	82	115	0	0

Threshold	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable	State	Min case	Median case	Max case	Number of geos	Percent suppressed	Percent unstable
Census tract		0.004	4	12	1391	88	100							
Total, 5k		0.2	7	30	706	41	87							
Total, 10k		3	13	46	360	3	35							
Total, 15k		2	19	54	252	0.8	7							
Total, 20k	***Missouri***	**9**	**25**	**71**	**191**	**0**	**1**							
Total, 25k		15	31	78	151	0	0							
Sub, 1k		3	8	32	571	21	78							
Sub, 2.5k		4	19	56	252	0.4	6							
Sub, 5k		19	38	97	127	0	0							

Note: All case counts shown are expected census tract-level case counts. Any numbers greater than 1 have been rounded to the nearest whole number and are referred to in the text by the complete value. Total population 20,000 persons is the recommended aggregation scheme for rarer outcomes. Sub-population aggregation thresholds were as follows: 1000 persons (333 persons 0–4 years old and 667 persons 65+ years old); 2500 persons (833 persons 0–4 years old and 1667 persons 65+ years old); and 5000 persons (1667 persons 0–4 years old and 3333 persons 65+ years old).

Table 4

Overview of aggregation schemes and the recommended median case count ranges and population thresholds (for a single year of data).

Aggregation scheme	Median case count range	Population threshold
Census tract	≥17.0 cases	Census tract *
Common outcome	7.3 to 16.9 cases	Total population 5000 persons
Rare outcome	1.9 to 7.2 cases	Total population 20,000 persons

* *Note:* The census tract population threshold is based on the fact that census tracts are built with an average population of 4000 persons (with a wide range).