BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-033373
Article Type:	Research
Date Submitted by the Author:	02-Aug-2019

Complete List of Authors:	Lee, David; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine; New York University School of Medicine, Department of Population Health Feldman, Justin; New York University School of Medicine, Department of Population Health Osorio, Marcela; New York University School of Medicine, Koziatek, Christian; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nguyen, Michael; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nagappan, Ashwini; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Shim, Christopher; California Northstate University College of Medicine Vinson, Andrew; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Thorpe, Lorna; New York University School of Medicine, Department of Population Health McGraw, Nancy; Sullivan County Public Health Services

Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, HEALTH SERVICES ADMINISTRATION & MANAGEMENT

Title: Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data

Authors
David C. Lee, MD, MS[1,2]
Justin M. Feldman, MPH, ScD[2]
Marcela Osorio[1]
Christian A. Koziatek, MD[1]
Michael V. Nguyen[1]
Ashwini Nagappan[1]
Christopher J. Shim, JD, MS[3]
Andrew J. Vinson, MA, GISP[1]
Lorna E. Thorpe, PhD, MPH[2]
Nancy A. McGraw, LCSW, MBA, MPH[4]

Author Affiliations
1. Ronald O. Perelman Department of Emergency Medicine, New York University School of Medicine, New York, NY, USA
2. Department of Population Health, New York University School of Medicine, New York, NY, USA
3. California Northstate University College of Medicine, Elk Grove, CA, USA
4. Sullivan County Public Health Department, Liberty, NY, USA

Corresponding Author
David C. Lee, MD, MS
227 East 30th Street, Room 107
New York, NY 10016
Phone: (212) 562-6561
Email: David.Lee@nyulangone.org

Manuscript Word count: 2,995

ABSTRACT

Objectives: Some of the most pressing health problems are found in rural America. However, the surveillance needed to track and prevent disease in these regions is severely lacking. The objective of our study was to perform a comprehensive health survey of a single rural county to assess the validity of using emergency claims data to estimate rural disease prevalence at a sub-county level.

Design: We compared chronic disease prevalence estimates using emergency department (ED) claims data versus a mailed health survey designed to capture a substantial proportion of residents in New York's rural Sullivan County in 2017-2018. We compared age- and gender-adjusted prevalence of hypertension, hyperlipidemia, diabetes, cancer, asthma, and COPD/emphysema among nine sub-county areas.

Results: Our countywide mailed survey obtained 6,675 completed responses for a response rate of 30.4%. This sample represented more than 12% of the estimated 53,020 adults in Sullivan County. Using emergency claims data, we identified 34,576 adults from Sullivan County who visited an emergency department at least once during 2011-2015. At a sub-county level, prevalence estimates from mailed surveys and emergency claims data correlated especially well for diabetes (r=0.90) and asthma (r=0.85). Using the larger sample from emergency claims data, we created geographically detailed maps of disease prevalence using geocoded addresses.

Conclusions: For select conditions, emergency claims data may be useful for tracking disease prevalence in rural areas and provide more geographically detailed estimates. For rural regions

lacking robust health surveillance, emergency claims data can inform how to geographically target efforts to prevent chronic disease.

Article Summary

Strengths and limitations of this study

- Validates the use of emergency claims data to perform geographically detailed surveillance in rural settings.
- Provides a standard for estimating disease prevalence at a local level by performing a countywide mailed survey.
- Limited by the accuracy of diagnosis codes found in claims data and is more accurate for conditions likely to be captured during emergency visits.
- Has the potential to improve rural health surveillance by using existing data to track the burden of chronic diseases.

INTRODUCTION

In New York State, Sullivan County has been ranked 61st out of 62 by the County Health Rankings just behind Bronx County in New York City.(1) Like many rural areas of America, Sullivan County has faced significant economic challenges, along with disparities in healthcare access.(2, 3) Though some of the most pressing health problems can be found in rural America, their public health institutions lack timely data needed to provide geographically detailed chronic disease surveillance.(4, 5) Nationwide health surveys, such as the Behavioral Risk Factor Surveillance System (BRFSS), often have inadequate coverage of these rural regions, and efforts to use models to extrapolate estimates of disease prevalence have questionable validity.(6, 7)

In recent years, there has been increasing interest in using alternative sources of data to track chronic disease prevalence.(8-10) Approaches using claims data and electronic health records have emerged among the potential options.(11) These data are collected routinely by state agencies and may provide a cost-effective, ready-to-analyze alternative to expensive and time-intensive traditional survey methods.(12) For instance, 1 in 5 Americans report having visited an emergency department (ED) in the past year, which provides a 20% population sample with a single year of data.(13, 14) However, these approaches need to be validated before widespread dissemination because, unlike surveys, they are not random population samples and may therefore not be representative.

There are several challenges that make estimation of chronic disease prevalence in rural areas difficult. First, few traditional health surveys have been performed in these areas with sample sizes adequate for sub-county level area estimation.(15, 16) In addition, there is similarly sparse data on the sociodemographic composition of these rural regions (as evidenced by wide confidence intervals for sub-county estimates of race and ethnicity). Furthermore, ZIP codes, county borders, and other geographic units are less likely to align in rural areas, limiting the possibility of attributing aggregated data to specific regions.(17) Thus, certain traditional survey techniques used to refine estimates based on underlying demographic characteristics (e.g., statistical weighting, adjustment, or stratification) become difficult to implement in rural areas.(18)

The goal of this study was to perform a comprehensive health survey of a single rural county in the United States. We report the results of a geographically distributed health survey delivered by mail to households within Sullivan County, New York. We then compare the disease prevalence estimates obtained from these surveys to a novel method that uses emergency claims data to identify areas with a higher burden of chronic disease.(14)

METHODS

Study Design

We administered a brief health survey by mail throughout Sullivan County during Fall 2017 and Spring 2018 to a random sample of residential addresses. We used survey data to estimate the age- and gender-standardized prevalence of several chronic diseases at a sub-county level. We also estimated disease prevalence using the comprehensive, all-payer New York Statewide Planning and Research Cooperative System (SPARCS) claims database. Our alternative measure was the proportion of ED patients with ≥ 1 diagnostic code for a given disease on ≥ 1 emergency visit during the period 2011-2015.(14) In each method, residents with an address located at a nursing or correctional facility were excluded to estimate prevalence for the non-institutionalized population. This study was approved by NYU School of Medicine's Institutional Review Board.

Mailed Health Surveys

To generate a sampling frame for our mailed health survey, we obtained point and parcel data for all mailing addresses in Sullivan County from the New York State GIS Clearinghouse (www.gis.ny.gov).(19) This data source was selected because it contained property class and land use data. Addresses were filtered to include any residential listing not marked as seasonal housing. We also included commercial addresses listed as apartments. This list of mailing addresses was then refined using an address verification service (www.smartystreets.com) to select valid, non-vacant mailable addresses.(20) As a substantial proportion of residents do not receive delivered mail in Sullivan County, we also queried the address verification service to find all valid, mailable PO boxes in the county. The final sampling frame consisted of 39,084 households located across 56 ZIP codes within Sullivan County.

Given the sparse population in some areas of the rural county, less-populated ZIP codes were oversampled to maximize the geographic coverage of the survey over the entire county. Using a quota sampling strategy, we mailed surveys to a random sample of 750 households for each ZIP code in our sampling frame. In ZIP codes with fewer than 750 households, all households were mailed a survey. Each health survey consisted of questions that first confirmed residence within Sullivan County and then asked a brief selection of health and demographic questions derived from the BRFSS.(21) Survey respondents were offered a $10 gift card for participation and a stamped return envelope was enclosed in the surveys. The Sullivan County Public Health Department also made local news outlets aware of the survey and fielded phone calls from local residents to confirm that the survey was legitimate.

Emergency Claims Data

Using the SPARCS, all-payer claims database, we identified all adult patients who had visited an ED located at a general acute care hospital in New York State between 2011 and 2015. We included all patients with a PO box or home address located within the borders of Sullivan County. Patients with more than one ED visit either at the same hospital or different hospitals were counted as a single observation by collapsing multiple visits using unique identifiers from SPARCS. The result was a listing of unique Sullivan County residents who had accessed emergency care at least once during the five-year period.

Study Outcomes

Our primary outcome was the prevalence of chronic disease at a sub-county level as identified by our mailed health survey or estimated using emergency claims data. In our mailed survey, respondents were asked if they had ever been diagnosed with hypertension, hyperlipidemia, cancer, diabetes, asthma, chronic obstructive pulmonary disease (COPD) or emphysema. In our analysis of emergency claims data, all available primary and secondary diagnosis codes across visits were scanned by individual for the presence of ≥ 1 diagnosis code during ≥ 1 ED visit for these same conditions. The codes from the International Classification of Diseases (ICD-9 and ICD-10) used were: hypertension (401-405 or I10-I16), hyperlipidemia (272 or E78), diabetes (250 or E10-E11), cancer (140-239 or C00-C96), asthma (493 or J45), and COPD/emphysema (491-492 or J43-J44). Thus, prevalence was estimated as a proportion: the number of unique ED patients with each of the listed conditions, divided by the total number of unique ED patients.

Statistical Analysis

To generate the sub-county areas in our analysis, we first grouped ZIP codes based on the Census-defined subdivisions (i.e., town borders) within Sullivan County. ZIP codes were assigned to these subdivisions based on the largest area of overlap given that ZIP code boundaries do not exactly align with town borders.(17) After grouping ZIP codes into these fifteen subdivisions, it was found that ten of these subdivisions had less than 2,000 households that received a mailed survey and were thus unlikely to obtain the minimum 500 survey responses, a benchmark set by the Centers for Disease Control and Prevention for obtaining acceptably narrow confidence intervals for prevalence estimation (Supplemental Figure 1).(18) Therefore, these less populated subdivisions were systematically merged with each other to form four sub-county areas with a sufficient number of sampled households. The result was nine sub-county areas made of five subdivisions with adequate sampling and four areas combining neighboring subdivisions to attain adequate sampling. (See Supplemental Table 1 for more details of aggregating ZIP codes into subdivisions and then sub-county areas).

In aggregating prevalence estimates between ZIP codes to create the sub-county areas, we used two weighting approaches. For the mailed survey, we applied design weights (the inverse probability of selection from the sampling frame) to account for our oversampling of less-populated ZIP codes. For the emergency claims data, we weighted ZIP code prevalence estimates by the inverse of the total number of unique ED patients divided by the Census estimate of adults aged 25 years and older for each ZIP code in Sullivan County to account for known differences in ED utilization based on proximity to the nearest hospital.(22) Prevalence estimates using both methods were then standardized to the overall age and gender distributions in Sullivan County from the most recent five-year 2012-2016 American Community Survey (ACS).(23) We then calculated Pearson correlation coefficients comparing the prevalence estimates obtained using the two methods at the sub-county level.

Geographic Analysis

We also performed geographically detailed surveillance using the larger sample of Sullivan County residents identified in emergency claims data. For the subset of patients with a geocodable home address, we calculated unadjusted disease prevalence among their 100 nearest neighbors identified in the population of unique ED patients. We then interpolated rasters from this point data using the inverse squared distance technique. Chronic disease prevalence maps were generated from these unadjusted prevalence estimations for diabetes, asthma, and hypertension, with categories based on standard deviations from the mean. For comparison, these maps were also created based on the 200 nearest neighbors to assess the influence of changing this parameter.

Statistical analyses were performed using Stata 14.2 (Statacorp; College Station, TX, 2015). Geographic analysis and mapping were performed using ArcGIS Desktop 10.5.1 (ESRI; Redlands, CA, 2017).

RESULTS

Mailed Survey Responses

Of the 24,141 surveys that we mailed to addresses within Sullivan County, approximately 20% were returned to sender even after using an address verification service. Of the 7,241 survey responses received, 216 were missing key demographic information or were otherwise incomplete, 248 were not residents of Sullivan County, and 22 were located at a nursing or correctional facility. In addition, only 80 respondents were aged 18 to 24 years old, which we deemed too few for inclusion in the study. Therefore, we limited study results to adults aged 25 years and older. Using the AAPOR RR2 definition for mail surveys of unnamed persons, our response rate was 30.4%.(24)

Population Characteristics

The countywide mailed survey received valid responses for 6,675 adults or 12.6% of the adult population 25 years and older in Sullivan County. Using five years of emergency claims data, we were able to identify 65.2% of the adult population 25 years and older in Sullivan County. In comparison to ACS 2012-2016 Census estimates, survey respondents were notably older (42.5% versus Census estimate of 23.9% aged 65 years and older). In comparison, the population of unique ED patients was slightly younger (39.2% versus Census estimate of 33.0% aged 25 to 44 years old). A higher proportion of survey respondents were female (60.7% versus Census estimate of 49.2%). Also, a higher proportion of survey respondents were non-Hispanic white (88.7% versus Census estimate of 73.0%). However, the sex and race/ethnicity distributions of the unique ED patient population were similar to Census estimates (Table 1).

Table 1: Demographic Comparisons Among Census Estimates and Data Sources

Demographic Comparisons	2012-2016 Census Estimates	County Wide Mailed Survey	Emergency Claims Data
Total	53,020	6,675	34,567
Age			
25 to 44	33.0%	15.9% (15.0% - 16.8%)	39.2% (38.7% - 39.7%)
45 to 64	43.1%	41.6% (40.4% - 42.7%)	36.9% (36.4% - 37.5%)
65 and older	23.9%	42.5% (41.3% - 43.7%)	23.9% (23.4% - 24.3%)
Sex			
Male	50.8%	39.3% (38.1% - 40.4%)	49.5% (48.9% - 50.0%)
Female	49.2%	60.7% (59.6% - 61.9%)	50.5% (50.0% - 51.1%)
Race / Ethnicity			
White	73.0%	88.7% (87.4% - 89.0%)	75.4% (74.9% - 75.8%)
Black	7.7%	2.3% (1.9% - 2.6%)	7.0% (6.7% - 7.3%)
Hispanic	15.0%	5.3% (4.8% - 5.9%)	10.5% (10.1% - 10.8%)
Asian	1.6%	0.9% (0.7% - 1.2%)	0.5% (0.4% - 0.6%)
Other	2.7%	2.8% (2.4% - 3.2%)	6.6% (6.3% - 6.9%)

Prevalence Estimates

The county-wide prevalence estimates using emergency claims data was higher than the mailed survey for diabetes, but lower for asthma (Table 2). The correlation by sub-county area was very strong for these two conditions respectively at r = 0.90 (95% CI of 0.60 to 0.98) and r = 0.85 (95% CI of 0.44 to 0.97). (See Supplemental Figure 2 for graphs of correlations). For all other conditions, the county-wide prevalence estimates using emergency claims data was lower than the mailed survey. These correlations were graded across conditions: moderate for hypertension (r = 0.46, CI: -0.30 to 0.86) and COPD/emphysema (r = 0.42, CI: -0.34 to 0.85), and weak for cancer (r = 0.39, CI: -0.37 to 0.84) and hyperlipidemia (r = 0.23, CI: -0.51 to 0.78). We displayed maps of prevalence estimates for diabetes, asthma, and hypertension based on survey results for the sub-county areas analyzed in Figure 1.

Table 2: County Disease Prevalence and Correlation at a Sub-County Level

Chronic Disease	County Wide Mailed Survey	Emergency Claims Data	Correlation Coefficient
Diabetes	12.7% (11.5% - 13.8%)	14.7% (14.3% - 15.1%)	0.90 (0.60 to 0.98)
Asthma	15.6% (14.1% - 17.1%)	8.7% (8.4% - 9.0%)	0.85 (0.44 to 0.97)
Hypertension	38.3% (36.7% - 39.8%)	36.1% (35.6% - 36.6%)	0.46 (-0.30 to 0.86)
COPD / Emphysema	7.0% (6.2% - 7.9%)	4.0% (3.8% - 4.2%)	0.42 (-0.34 to 0.85)
Cancer	10.7%	3.9%	0.39

	(9.8% - 11.5%)	(3.7% - 4.2%)	(-0.37 to 0.84)
Hyperlipidemia	33.9% (32.2% - 35.5%)	21.2% (20.8% - 21.7%)	0.23 (-0.51 to 0.78)

Emergency Department Surveillance

Among the 34,567 unique patients identified from emergency claims data, 76% had a geocodable home address, 20% were PO box only, and 4% were not geocodable but had a ZIP code located fully within Sullivan County. Using the 100 nearest neighbors among patients with a geocodable home address, we estimated unadjusted prevalence at the geocoded location of each patient and created interpolated rasters to provide a more geographically detailed maps of diabetes, asthma, and hypertension prevalence (Figure 2). These maps were able to identify localized clusters of disease throughout the county with greater geographic detail.

DISCUSSION

The intensity of health problems experienced by residents living in rural areas of the country underscores the need for improving our methods of health surveillance.(2, 3) Our study's findings demonstrate a novel solution that uses emergency claims data to estimate chronic disease prevalence at a sub-county level. These estimates are important for identifying key hotspots of disease, which may reveal previously unexplored risk factors that increase disease burden in rural America and guide efforts to prevent chronic disease in specific geographic areas that experience the worst health outcomes.(25) Current health surveillance techniques rely on traditional methods such as telephone-based surveys. Not only are these methods costly and time-intensive, but due to dramatic shifts in phone use, response rates over the past two decades have dropped dramatically from around 36% to 9%.(26, 27) The sample size of a large national health survey such as the BRFSS is inadequate for generating precise estimates of disease prevalence even at the county level for much of rural America.(28)

Recent efforts to provide greater geographic coverage have focused on approaches that use the data in adequately sampled areas and statistical models to extrapolate disease estimates for poorly sampled areas largely based on sociodemographic factors.(15) But many of these techniques have not been validated, and in the few instances when they have been compared, these approaches do not always work as well as expected.(6, 7) Our mailed health survey found that countywide adjusted diabetes prevalence was 12.7%. This estimate is much higher than the CDC's most recent estimate of 9.5% in 2015, which is based on a modelling approach. For a given area, these modelling approaches can be especially imprecise when used to estimate disease prevalence in areas with low response rates, which includes many rural regions. Other efforts to advance health surveillance methods have experimented with the use of claims data and electronic health records to provide estimates of disease prevalence. A recent study demonstrated that emergency claims data could be used to estimate chronic disease prevalence in New York City, and this approach was validated with results obtained from an annually performed citywide health survey. In this urban study, it was found that conditions including diabetes, hypertension, and asthma had correlations of 0.86, 0.88, and 0.77 respectively when analyzed among 34 sub-county areas.(14)

With our novel method of using emergency claims data to estimate chronic disease burden, we identified health records for a substantial majority of all adults in Sullivan County using five years of emergency claims data. Furthermore, the demographic patterns among this population of unique ED patients were much closer to Census estimates than our countywide mailed survey. Under-representation of certain demographic groups, especially minorities, is a common problem of traditional survey methods that can be adjusted for, as long as geographically matched sociodemographic data exists.(29) In rural areas where Census estimates for race and ethnicity often have wide confidence intervals, emergency claims data may provide an alternative population sample that closely mirrors the underlying population in a given region. For some conditions such as diabetes and asthma, we found strong correlation between the two estimation methods for sub-county disease prevalence. For the other conditions studied, the strength of correlation was weaker. This may be attributable to disease-specific differences in the validity of both ED claims data and self-reported survey data. Prior research has shown that, for both data sources, validity is routinely higher for diabetes and asthma, but lower for other conditions like hyperlipidemia, with low sensitivity (i.e., under-diagnosis) being the reason for poor correlation.(8-10)

Emergency claims data are already widely collected around the country, can capture a large population sample, and in some areas include address data that can be used to precisely identify where patients live. By geocoding these addresses, more precise health surveillance can provide detailed maps of disease burden. This granular level of geographic detail is important because localized hot spots of disease might otherwise be hidden as they are averaged out by neighboring areas of low disease prevalence. However, some important caveats should be understood before employing these methods. There is some variation in how accurately some hospitals capture chronic disease conditions using diagnosis codes.(30) In addition, for some parts of rural America, mail is only delivered to PO boxes, therefore the more geographically detailed maps of disease prevalence based on geocoded data may not be accurate in these regions where mail is not delivered directly.

Limitations

Because surveys did not ask respondents to report household size, single-adult households are likely overrepresented. However, we age- and gender-standardized rates to the overall population in Sullivan County, which may partially reduce this bias. While our adjustment methods reduced age- and gender-specific non-response bias, we were unable to standardize by race and ethnicity due to the very small proportion of minorities in several areas of Sullivan County. Given that minorities often have higher rates of chronic disease and tend to have lower response rates, our mailed survey may have underestimated disease prevalence. Our method that used emergency claims data to estimate disease prevalence is subject to many of the limitations associated with the use of administrative data. Fidelity of coding some variables including race, ethnicity, and diagnosis codes can vary by hospital and may impact resulting disease prevalence estimates. Also, these claims data are often available about a year after they have been filed, thus there is some lag in reporting.

CONCLUSION

We found that for select conditions, ED data may be useful for tracking disease prevalence in rural areas and may provide more geographically precise estimates. Given the infrastructure already in place to collect this data, efforts could be focused on collecting more accurate diagnosis codes and more detailed geographic data. This approach could potentially help geographically target efforts to prevent chronic disease.

Acknowledgments

The content of the study reflects the views of the authors and not the official position of the Sullivan County Public Health Department nor the NYU School of Medicine.

Author contributions

Study conception and design: DCL, LET, NAM; acquisition of the data: DCL, MO, MVN, AN, AJV; analysis and interpretation of the data: DCL, JMF, MO, CAK, CJS, AJV, LET, MAN; drafting of the manuscript: DCL; critical revision of the manuscript for intellectual content: JMF, MO, CAK, MVN, AN, CJS, AJV, LET, MAN. DCL is the guarantor of this work and had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This work was supported by the New York State Health Foundation grant number 16-04083.

Competing of interests

None declared.

Patient consent for publication

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination of our research.

Ethics approval

This study was approved by NYU School of Medicine's Institutional Review Board.

Data availability statement

Deidentified participant data from the countywide health survey from 2017-2018 can be made available upon reasonable request by contacting Dr. David Lee at david.lee@nyumc.org so long as the requester agrees to the following conditions: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

REFERENCES

1. Roadmaps CHRa. 2018 County Health Rankings Report. Madison, WI: University of Wisconsin Population Health Institute; 2018 [cited 2018 November 8th]; Available from: http://www.countyhealthrankings.org/sites/default/files/state/downloads/CHR2018_NY.pdf.

2. Harris JK, Beatty K, Leider JP, Knudson A, Anderson BL, Meit M. The Double Disparity Facing Rural Local Health Departments. Annual review of public health. 2016;37:167-84.

3. Brundisini F, Giacomini M, DeJean D, Vanstone M, Winsor S, Smith A. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. Ontario health technology assessment series. 2013;13(15):1-33.

4. Puma JE, Belansky ES, Garcia R, Scarbro S, Williford D, Marshall JA. A Community-Engaged Approach to Collecting Rural Health Surveillance Data. The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association. 2017 Jun;33(3):257-65.

5. Bellamy GR, Bolin JN, Gamm LD. Rural Healthy People 2010, 2020, and beyond: the need goes on. Family & community health. 2011 Apr-Jun;34(2):182-8.

6. Feldman JM. Re: Obesity and place: Chronic disease in the 500 largest U.S. cities. Obesity research & clinical practice. 2018 Sep - Oct;12(5):490.

7. Wang Y, Holt JB, Zhang X, Lu H, Shah SN, Dooley DP, et al. Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013. Preventing chronic disease. 2017 Oct 19;14:E99.

8. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying persons with diabetes using Medicare claims data. American journal of medical quality : the official journal of the American College of Medical Quality. 1999 Nov-Dec;14(6):270-7.

9. Tessier-Sherman B, Galusha D, Taiwo OA, Cantley L, Slade MD, Kirsche SR, et al. Further validation that claims data are a useful tool for epidemiologic research on hypertension. BMC public health. 2013 Jan 18;13:51.

10. Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease. Comparing administrative data and self-reports. Medical care. 1997 Sep;35(9):932-47.

11. Colantonio LD, Levitan EB, Yun H, Kilgore ML, Rhodes JD, Howard G, et al. Use of Medicare Claims Data for the Identification of Myocardial Infarction: The Reasons for Geographic And Racial Differences in Stroke (REGARDS) Study. Medical care. 2018 Oct 22.

12. Hope KG, Merritt TD, Durrheim DN, Massey PD, Kohlhagen JK, Todd KW, et al. Evaluating the utility of emergency department syndromic surveillance for a regional public health service. Communicable diseases intelligence quarterly report. 2010 Sep;34(3):310-8.

13. Centers for Disease Control and Prevention. Health, United States, 2013: With Special Feature on Prescription Drugs. Hyattsville, MD: National Center for Health Statistics, 2014.

14. Lee DC, Long JA, Wall SP, Carr BG, Satchell SN, Braithwaite RS, et al. Determining Chronic Disease Prevalence in Local Populations Using Emergency Department Surveillance. Am J Public Health. 2015 Sep;105(9):e67-74.

15. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. American journal of preventive medicine. 2004 Jun;26(5):453-60.

16. Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. American journal of epidemiology. 2015 Jul 15;182(2):127-37.

17. Coburn AF, MacKinney AC, McBride TD, Mueller KJ, Slifkin RT, Wakefield MK. Choosing Rural Definitions: Implications for Health Policy Iowa City, IA: Rural Policy Research Institute Health Panel, 2007.

18. Centers for Disease Control and Prevention. SMART BRFSS MMSA Methodology. Atlanta, GA2016 [cited 2018 November 8th]; Available from: https://www.cdc.gov/brfss/smart/2016/2016_SMART_BRFSS_MMSA_Methodology.pdf.

19. NYS ITS GIS Program Office. NYS GIS Clearinghouse. Albany, NY [cited 2018 November 8th]; Available from: www.gis.ny.gov

20. Mullen PD, Savas LS, Bundy LT, Haardorfer R, Hovell M, Fernandez ME, et al. Minimal intervention delivered by 2-1-1 information and referral specialists promotes smoke-free homes among 2-1-1 callers: a Texas generalisation trial. Tobacco control. 2016 Oct;25(Suppl 1):i10-i8.

21. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. BMC medical research methodology. 2013 Mar 24;13:49.

22. Lee DC, Doran KM, Polsky D, Cordova E, Carr BG. Geographic variation in the demand for emergency care: A local population-level analysis. Healthcare (Amsterdam, Netherlands). 2016 Jun;4(2):98-103.

23. Klein RJ, Schoenborn CA. Age adjustment using the 2000 projected U.S. population. Healthy People Statistical Notes. Hyattsville, Maryland: National Center for Health Statistics; January 2001.

24. American Association for Public Opinion Research. Mailed Surveys of Unnamed Persons. Oakbrook Terrace, ILApril 2016 [cited 2018 November 8th]; Available from: https://www.aapor.org/getattachment/Standards-Ethics/Standard-Definitions-(1)/Mail_Surveys_Unnamed_Persons_042216.pdf.aspx.

25. Brownson RC, Fielding JE, Green LW. Building Capacity for Evidence-Based Public Health: Reconciling the Pulls of Practice and the Push of Research. Annual review of public health. 2018 Apr 1;39:27-53.

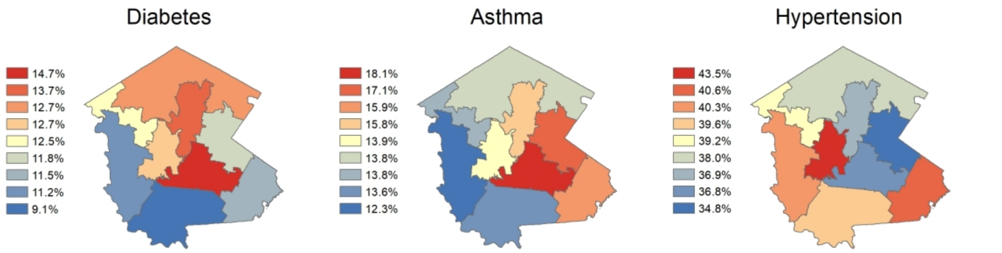
26. Galea S, Tracy M. Participation rates in epidemiologic studies. Annals of epidemiology. 2007 Sep;17(9):643-53.

27. Keeter S, Hatley N, Kennedy C, Lau A. What Low Response Rates Mean for Telephone Surveys. Washington DC: Pew Research Center; 2017 [cited 2017 July 25th]; Available from: http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/.

28. O'Connor A, Wellenius G. Rural-urban disparities in the prevalence of diabetes and coronary heart disease. Public health. 2012 Oct;126(10):813-20.

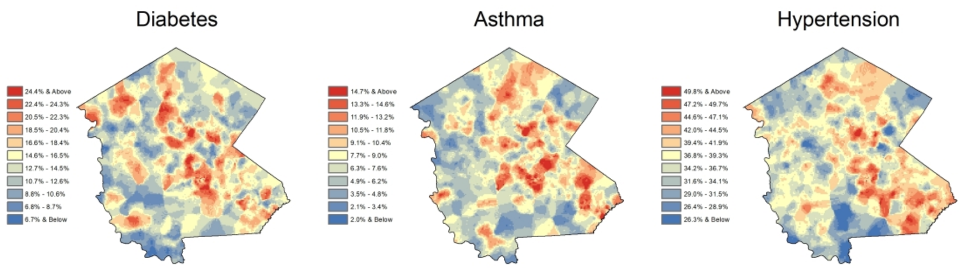
29. Link MW, Mokdad AH, Stackhouse HF, Flowers NT. Race, ethnicity, and linguistic isolation as determinants of participation in public health surveillance surveys. Preventing chronic disease. 2006 Jan;3(1):A09.

30. Lee DC, Swartz JL, Koziatek CA, Vinson AJ, Athens JK, Yi SS. Assessing the Reliability of Performing Citywide Chronic Disease Surveillance Using Emergency Department Data from Sentinel Hospitals. Population health management. 2017 Mar 24.



Sub-County Estimates of Adjusted Disease Prevalence Based on Mailed Survey Responses

228x76mm (300 x 300 DPI)



Geographically Detailed Estimates of Unadjusted Disease Prevalence Based on Emergency Claims Data

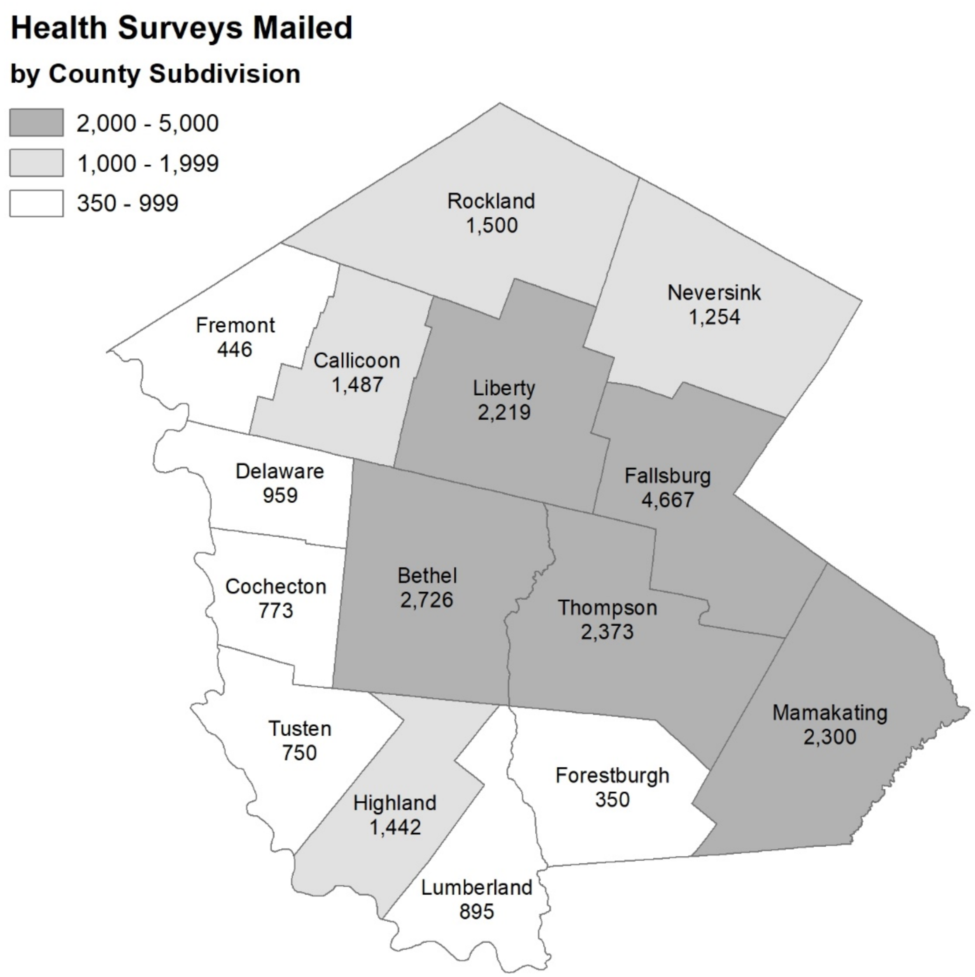
304x101mm (300 x 300 DPI)

Supplemental Table 1: Households, Mailings, and Responses by Geographic Level in Sullivan County

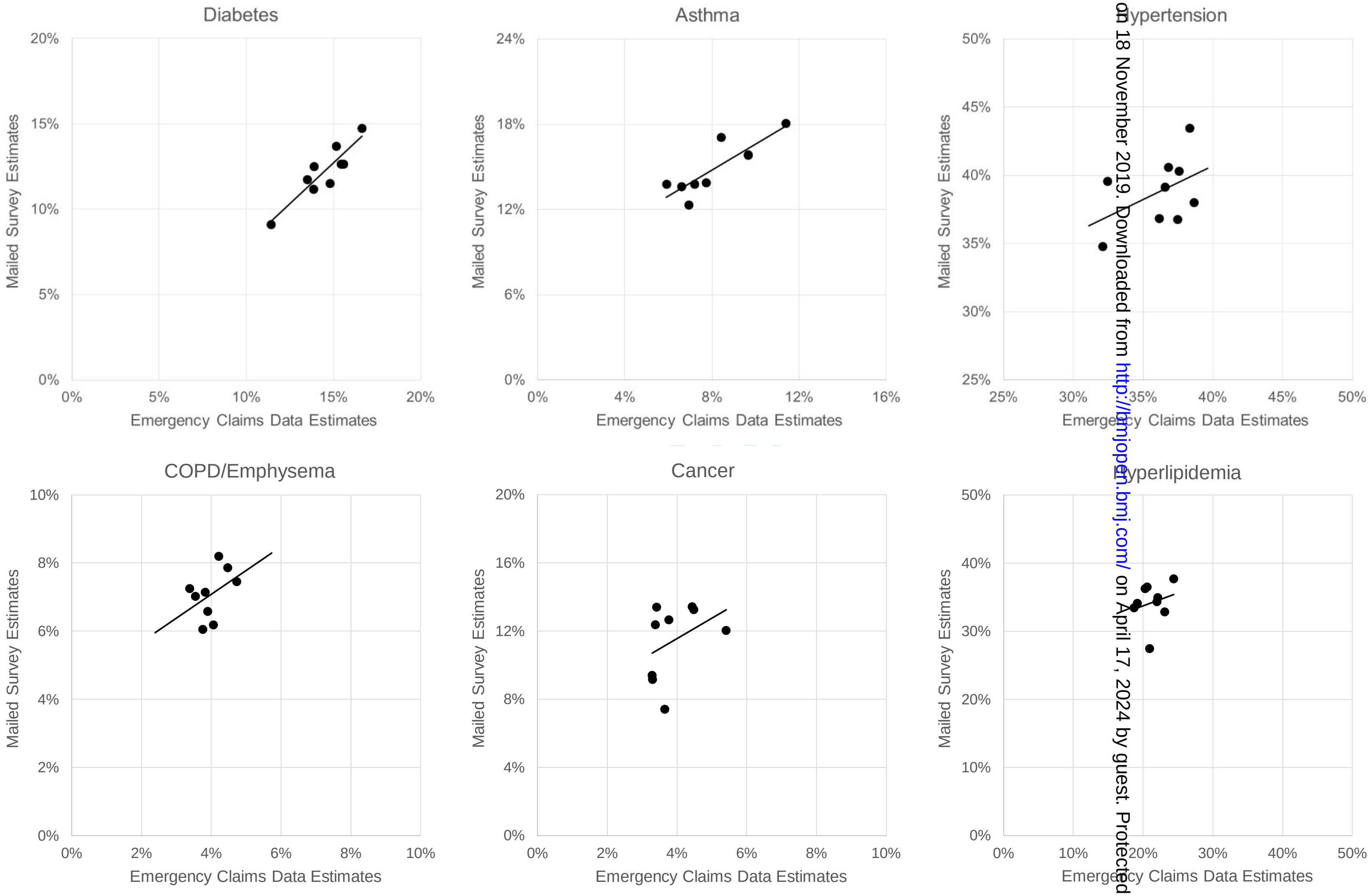
Zip Code	Population	Households	Mailings	Responses	Subdivison	Population	Households	Mailings	Responses	Subcounty	Population	Households	Mailings	Responses
12720	143	159	159	51	Bethel	2,900	2,774	2,726	719	Bethel	2,900	2,774	2,726	719
12749	266	382	382	97	Bethel					Bethel				
12762	288	323	323	104	Bethel					Bethel				
12778	437	618	618	179	Bethel					Bethel				
12783	1,396	798	750	157	Bethel					Bethel				
12786	370	494	494	131	Bethel					Bethel				
12724	217	199	199	85	Callicoon	2,175	1,731	1,487	526	Callicoon / Fremont	2,817	2,177	1,933	707
12748	1,172	994	750	205	Callicoon					Callicoon / Fremont				
12766	309	179	179	68	Callicoon					Callicoon / Fremont				
12791	477	359	359	168	Callicoon					Callicoon / Fremont				
12736	89	84	84	27	Fremont	642	446	446	181	Callicoon / Fremont				
12741	157	159	159	69	Fremont					Callicoon / Fremont				
12760	353	152	152	56	Fremont					Callicoon / Fremont				
12767	43	51	51	29	Fremont					Callicoon / Fremont				
12726	959	551	551	186	Cochecton	1,142	773	773	248	Cochecton / Delaware / Tusten	4,139	2,842	2,482	763
12752	183	222	222	62	Cochecton					Cochecton / Delaware / Tusten				
12723	1,391	893	750	212	Delaware	1,711	1,102	959	289	Cochecton / Delaware / Tusten				
12745	127	92	92	36	Delaware					Cochecton / Delaware / Tusten				
12750	193	117	117	41	Delaware					Cochecton / Delaware / Tusten				
12764	1,286	967	750	226	Tusten	1,286	967	750	226	Cochecton / Delaware / Tusten				
12733	1,070	550	550	79	Fallsburg	8,849	5,737	4,667	966	Fallsburg	8,849	5,737	4,667	966
12738	327	132	132	34	Fallsburg					Fallsburg				
12747	1,143	735	735	170	Fallsburg					Fallsburg				
12759	886	574	574	132	Fallsburg					Fallsburg				
12763	409	426	426	129	Fallsburg					Fallsburg				
12779	1,244	1,162	750	114	Fallsburg					Fallsburg				
12788	2,381	1,005	750	180	Fallsburg					Fallsburg				
12789	1,389	1,153	750	128	Fallsburg					Fallsburg				
12729	82	41	41	16	Forestburgh	781	350	350	122	Forestburgh /. Highland / Lumberland	4,314	2,763	2,687	856
12777	699	309	309	106	Forestburgh					Forestburgh /. Highland / Lumberland				
12719	1,042	520	520	179	Highland	2,013	1,442	1,442	493	Forestburgh /. Highland / Lumberland				
12732	577	480	480	155	Highland					Forestburgh /. Highland / Lumberland				
12743	242	184	184	66	Highland					Forestburgh /. Highland / Lumberland				
12792	152	258	258	93	Highland					Forestburgh /. Highland / Lumberland				
12737	1,382	826	750	193	Lumberland	1,520	971	895	241	Forestburgh /. Highland / Lumberland				
12770	138	145	145	48	Lumberland					Forestburgh /. Highland / Lumberland				
12734	697	652	652	162	Liberty	6,119	4,875	2,219	620	Liberty	6,119	4,875	2,219	620
12754	4,702	3,406	750	178	Liberty					Liberty				
12768	532	556	556	175	Liberty					Liberty				
12787	188	261	261	105	Liberty					Liberty				
12721	4,153	2,108	750	166	Mamakating	8,519	5,368	2,300	578	Mamakating	8,519	5,368	2,300	578
12722	77	200	200	59	Mamakating					Mamakating				
12769	156	100	100	40	Mamakating					Mamakating				
12781	211	133	133	42	Mamakating					Mamakating				
12785	704	367	367	63	Mamakating					Mamakating				
12790	3,218	2,460	750	208	Mamakating					Mamakating				
12725	143	103	103	24	Neversink	2,046	1,286	1,254	400	Neversink / Rockland	6,736	4,134	2,754	789
12740	1,208	782	750	236	Neversink					Neversink / Rockland				
12765	695	401	401	140	Neversink					Neversink / Rockland				
12758	2,966	1,997	750	204	Rockland	4,690	2,848	1,500	389	Neversink / Rockland				
12776	1,724	851	750	185	Rockland					Neversink / Rockland				
12701	7,795	6,044	750	228	Thompson	10,104	8,414	2,373	677	Thompson	10,104	8,414	2,373	677
12742	62	197	197	66	Thompson					Thompson				
12751	561	564	564	174	Thompson					Thompson				
12775	1,648	1,497	750	170	Thompson					Thompson				
12784	38	112	112	39	Thompson					Thompson				
Total	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**

Note: Population estimates are based on ZCTAs and may not align with ZIP codes. A population adjustment was required in ZIP code 12729 due to gross mismatch between ACS estimate and number of households from sampling frame.

Supplemental Figure 1: Health Survey Mailed by County Subdivision



Supplemental Figure 2: Comparison of Sub-County Estimates of Disease Prevalence



Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data: A Cross-Sectional Comparison of Survey versus Claims Data in Sullivan County, New York

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-033373.R1
Article Type:	Original research
Date Submitted by the Author:	27-Sep-2019
Complete List of Authors:	Lee, David; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine; New York University School of Medicine, Department of Population Health Feldman, Justin; New York University School of Medicine, Department of Population Health Osorio, Marcela; New York University School of Medicine, Koziatek, Christian; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nguyen, Michael; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nagappan, Ashwini; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Shim, Christopher; California Northstate University College of Medicine Vinson, Andrew; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Thorpe, Lorna; New York University School of Medicine, Department of Population Health McGraw, Nancy; Sullivan County Public Health Services
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Diabetes and endocrinology, Emergency medicine, Epidemiology, Public health, Research methods
Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™
Manuscripts

Title: Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data: A Cross-Sectional Comparison of Survey versus Claims Data in Sullivan County, New York

Authors
David C. Lee, MD, MS[1,2]
Justin M. Feldman, MPH, ScD[2]
Marcela Osorio[1]
Christian A. Koziatek, MD[1]
Michael V. Nguyen[1]
Ashwini Nagappan[1]
Christopher J. Shim, JD, MS[3]
Andrew J. Vinson, MA, GISP[1]
Lorna E. Thorpe, PhD, MPH[2]
Nancy A. McGraw, LCSW, MBA, MPH[4]

Author Affiliations
1. Ronald O. Perelman Department of Emergency Medicine, New York University School of Medicine, New York, NY, USA
2. Department of Population Health, New York University School of Medicine, New York, NY, USA
3. California Northstate University College of Medicine, Elk Grove, CA, USA
4. Sullivan County Public Health Department, Liberty, NY, USA

Corresponding Author
David C. Lee, MD, MS
227 East 30th Street, Room 107
New York, NY 10016
Phone: (212) 562-6561
Email: David.Lee@nyulangone.org

Manuscript Word count: 3,372

ABSTRACT

Objectives: Some of the most pressing health problems are found in rural America. However, the surveillance needed to track and prevent disease in these regions is lacking. Our objective was to perform a comprehensive health survey of a single rural county to assess the validity of using emergency claims data to estimate rural disease prevalence at a sub-county level.

Design: We performed a cross-sectional study of chronic disease prevalence estimates using emergency department claims data versus mailed health surveys designed to capture a substantial proportion of residents in New York's rural Sullivan County.

Setting: Sullivan County, a rural county ranked second-to-last for health outcomes in New York State.

Participants: Adult residents of Sullivan County aged 25 years and older who responded to the health survey in 2017-2018 or had at least one emergency department visit in 2011-2015.

Outcome Measures: We compared age- and gender-adjusted prevalence of hypertension, hyperlipidemia, diabetes, cancer, asthma, and COPD/emphysema among nine sub-county areas.

Results: Our countywide mailed survey obtained 6,675 completed responses for a response rate of 30.4%. This sample represented more than 12% of the estimated 53,020 adults in Sullivan County. Using emergency claims data, we identified 34,576 adults from Sullivan County who visited an emergency department at least once during 2011-2015. At a sub-county level, prevalence estimates from mailed surveys and emergency claims data correlated especially well for diabetes (r=0.90) and asthma (r=0.85). Other conditions were not as well correlated (range: 0.23-0.46). Using emergency claims data, we created more geographically detailed maps of disease prevalence using geocoded addresses.

Conclusions: For select conditions, emergency claims data may be useful for tracking disease prevalence in rural areas and provide more geographically detailed estimates. For rural regions lacking robust health surveillance, emergency claims data can inform how to geographically target efforts to prevent chronic disease.

Article Summary

Strengths and limitations of this study

- Validates the use of emergency claims data to perform geographically detailed surveillance in rural settings.
- Provides a standard for estimating disease prevalence at a local level by performing a countywide mailed survey.
- Limited by the accuracy of diagnosis codes found in claims data and is more accurate for conditions likely to be captured during emergency visits.
- Has the potential to improve rural health surveillance by using existing data to track the burden of chronic diseases.

INTRODUCTION

In New York State, Sullivan County has been ranked 61st out of 62 by the County Health Rankings based on rates of premature death and quality of life (poor overall, physical, mental health and low birthweights) just behind Bronx County in New York City.[1] Located just two hours northwest of New York City, Sullivan County is rural and more than 70% of its residents are White. Like many rural areas of America, Sullivan County has faced significant economic challenges, along with disparities in healthcare access.[2 3] Though some of the most pressing health problems can be found in rural America, their public health institutions lack timely data needed to provide geographically detailed chronic disease surveillance.[4 5] Nationwide health surveys, such as the Behavioral Risk Factor Surveillance System (BRFSS), often have inadequate coverage of these rural regions, and efforts to use models to extrapolate estimates of disease prevalence have questionable validity.[6 7]

In recent years, there has been increasing interest in using alternative sources of data to track chronic disease prevalence.[8-10] Approaches using claims data and electronic health records have emerged among the potential options.[11] These data are collected routinely by state agencies and may provide a cost-effective, ready-to-analyze alternative to expensive and time-intensive traditional survey methods.[12] For instance, 1 in 5 Americans report having visited an emergency department (ED) in the past year, which provides a 20% population sample with a single year of data.[13 14] However, these approaches need to be validated before widespread dissemination because, unlike surveys, they are not random population samples and may therefore not be representative.

There are several challenges that make estimation of chronic disease prevalence in rural areas difficult. The Centers for Disease Control and Prevention (CDC) has started to use modelling approaches with Bayesian and spatial smoothing with BRFSS data to estimate county-level disease prevalence in rural areas.[15] But, few traditional health surveys have been performed in these areas with sample sizes adequate for sub-county level area estimation.[16 17] In addition, there is similarly sparse data on the sociodemographic composition of these rural regions (as evidenced by wide confidence intervals for sub-county estimates of race and ethnicity). Furthermore, ZIP codes, county borders, and other geographic units are less likely to align in rural areas, limiting the possibility of attributing aggregated data to specific regions.[18] In addition, certain traditional survey techniques used to refine estimates based on underlying demographic characteristics (e.g., statistical weighting, adjustment, or stratification) that are often performed with Census data cannot be readily applied in rural areas especially if there is insufficient data.[19]

The goal of this study was to perform a comprehensive health survey of a single rural county in the United States. We report the results of a geographically distributed health survey delivered by mail to households within Sullivan County, New York. We then compare the disease prevalence estimates obtained from these surveys to a novel method that uses emergency claims data to identify areas with a higher burden of chronic disease.[14]

METHODS

Study Design

We administered a brief health survey by mail throughout Sullivan County during Fall 2017 and Spring 2018 to a random sample of residential addresses. We used survey data to estimate the age- and gender-standardized prevalence of several chronic diseases at a sub-county level. We also estimated disease prevalence using the comprehensive, all-payer New York Statewide Planning and Research Cooperative System (SPARCS) claims database. Our alternative measure was the proportion of ED patients with ≥ 1 diagnostic code for a given disease on ≥ 1 emergency visit during the period 2011-2015.[14] In each method, residents with an address located at a nursing or correctional facility were excluded to estimate prevalence for the non-institutionalized population. This study was approved by NYU School of Medicine's Institutional Review Board.

Mailed Health Surveys

To generate a sampling frame for our mailed health survey, we obtained point and parcel data for all mailing addresses in Sullivan County from the New York State GIS Clearinghouse (www.gis.ny.gov).[20] This data source was selected because it contained property class and land use data. Addresses were filtered to include any residential listing not marked as seasonal housing. We also included commercial addresses listed as apartments. This list of mailing addresses was then refined using an address verification service (www.smartystreets.com) to select valid, non-vacant mailable addresses.[21] As a substantial proportion of residents do not receive delivered mail in Sullivan County, we also queried the address verification service to find all valid, mailable PO boxes in the county. The final sampling frame consisted of 39,084 households located across 56 ZIP codes within Sullivan County.

Given the sparse population in some areas of the rural county, less-populated ZIP codes were oversampled to maximize the geographic coverage of the survey over the entire county. To do so, we used a quota sampling strategy. We mailed surveys to a random sample of 750 households for each ZIP code in our sampling frame. In ZIP codes with fewer than 750 households, all households were mailed a survey. Each health survey consisted of questions that first confirmed residence within Sullivan County and age over 18 years, and then asked a brief selection of health and demographic questions derived from the BRFSS.[22] For households with multiple residents, we asked that only one adult respond to the survey. We mailed a survey to 24,141 or 62% of the households in our sampling frame. Survey respondents were offered a $10 gift card for participation and a stamped return envelope was enclosed in the surveys. The Sullivan County Public Health Department also made local news outlets aware of the survey and fielded phone calls from local residents to confirm that the survey was legitimate.

Emergency Claims Data

Using the SPARCS, all-payer claims database, we identified all adult patients who had visited an ED located at a general acute care hospital in New York State between 2011 and 2015. We included all patients with a PO box or home address located within the borders of Sullivan County. Patients with more than one ED visit either at the same hospital or different hospitals were counted as a single observation by collapsing multiple visits using unique identifiers from SPARCS. The result was a listing of unique Sullivan County residents who had accessed emergency care at least once during the five-year period.

Study Outcomes

Our primary outcome was the prevalence of chronic disease at a sub-county level as identified by our mailed health survey or estimated using emergency claims data. In our mailed survey, respondents were asked if they had ever been diagnosed with hypertension, hyperlipidemia, cancer, diabetes, asthma, chronic obstructive pulmonary disease (COPD) or emphysema. In our analysis of emergency claims data, all available primary and secondary diagnosis codes across visits were scanned by individual for the presence of ≥ 1 diagnosis code during ≥ 1 ED visit for these same conditions. The codes from the International Classification of Diseases (ICD-9 and ICD-10) used were: hypertension (401-405 or I10-I16), hyperlipidemia (272 or E78), diabetes (250 or E10-E11), cancer (140-239 or C00-C96), asthma (493 or J45), and COPD/emphysema (491-492 or J43-J44). Thus, prevalence was estimated as a proportion: the number of unique ED patients with each of the listed conditions, divided by the total number of unique ED patients.

Statistical Analysis

To generate the sub-county areas in our analysis, we first grouped ZIP codes based on the Census-defined subdivisions (i.e., town borders) within Sullivan County. ZIP codes were assigned to these subdivisions based on the largest area of overlap given that ZIP code boundaries do not exactly align with town borders.[18] After grouping ZIP codes into these fifteen subdivisions, it was found that ten of these subdivisions had less than 2,000 households that received a mailed survey and were thus unlikely to obtain the minimum 500 survey responses, a benchmark set by the CDC for obtaining acceptably narrow confidence intervals for prevalence estimation (Supplemental Figure 1).[19] Therefore, these less populated subdivisions were systematically merged with each other based on proximity and population density to form four sub-county areas with a sufficient number of sampled households. The result was nine sub-county areas made of five subdivisions with adequate sampling and four areas combining neighboring subdivisions to attain adequate sampling. (See Supplemental Table 1 for more details of aggregating ZIP codes into subdivisions and then sub-county areas).

In aggregating prevalence estimates between ZIP codes to create the sub-county areas, we used two weighting approaches. For the mailed survey, we applied design weights (the inverse probability of selection from the sampling frame) to account for our oversampling of less-populated ZIP codes. For the emergency claims data, we weighted ZIP code prevalence estimates by the inverse of the total number of unique ED patients divided by the Census estimate of adults aged 25 years and older for each ZIP code in Sullivan County to account for known differences in ED utilization based on proximity to the nearest hospital.[23] Prevalence estimates using both methods were then standardized to the overall age and gender distributions in Sullivan County from the most recent five-year 2012-2016 American Community Survey (ACS).[24] We then calculated Pearson correlation coefficients comparing the prevalence estimates obtained using the two methods at the sub-county level. By convention, the strength of correlation was graded as very strong (0.80 - 1.00), strong (0.60 - 0.79), moderate (0.40 - 0.59), weak (0.20 - 0.39), and very weak (0.00 - 0.19).

Geographic Analysis

We also performed geographically detailed surveillance using the larger sample of Sullivan County residents identified in emergency claims data. For the subset of patients with a geocodable home address, we calculated unadjusted disease prevalence among their 100 nearest neighbors identified in the population of unique ED patients. We then interpolated rasters from this point data using the inverse squared distance technique. Chronic disease prevalence maps were generated from these unadjusted prevalence estimations for diabetes, asthma, and hypertension, with categories based on standard deviations from the mean. For comparison, these maps were also created based on the 200 nearest neighbors to assess the influence of changing this parameter.

Statistical analyses were performed using Stata 14.2 (Statacorp; College Station, TX, 2015). Geographic analysis and mapping were performed using ArcGIS Desktop 10.5.1 (ESRI; Redlands, CA, 2017).

RESULTS

Mailed Survey Responses

Of the 24,141 surveys that we mailed to addresses within Sullivan County, approximately 20% were returned to sender even after using an address verification service (Supplemental Figure 2). Of the 7,241 survey responses received, 216 were missing key demographic information or were otherwise incomplete, 248 were not residents of Sullivan County, and 22 were located at a nursing or correctional facility. In addition, only 80 respondents were aged 18 to 24 years old, which we deemed too few for inclusion in the study. Therefore, we limited study results to adults aged 25 years and older. Using the AAPOR RR2 definition for mail surveys of unnamed persons, our response rate was 30.4%.[25]

Population Characteristics

The countywide mailed survey received valid responses for 6,675 adults or 12.6% of the adult population 25 years and older in Sullivan County. Using five years of emergency claims data, we were able to identify 65.2% of the Census-estimated adult population 25 years and older in Sullivan County. In comparison to ACS 2012-2016 Census estimates, survey respondents were notably older (42.5% versus Census estimate of 23.9% aged 65 years and older). In comparison, the population of unique ED patients was slightly younger (39.2% versus Census estimate of 33.0% aged 25 to 44 years old). A higher proportion of survey respondents were female (60.7% versus Census estimate of 49.2%). Also, a higher proportion of survey respondents were non-Hispanic white (88.7% versus Census estimate of 73.0%). However, the sex and race/ethnicity distributions of the unique ED patient population were similar to Census estimates (Table 1).

Table 1: Demographic Comparisons Among Census Estimates and Data Sources

Demographic Comparisons	2012-2016 Census Estimates	County Wide Mailed Survey	Emergency Claims Data
Total	53,020	6,675	34,567
Age			
25 to 44	33.0%	15.9% (15.0% - 16.8%)	39.2% (38.7% - 39.7%)
45 to 64	43.1%	41.6% (40.4% - 42.7%)	36.9% (36.4% - 37.5%)
65 and older	23.9%	42.5% (41.3% - 43.7%)	23.9% (23.4% - 24.3%)
Sex			
Male	50.8%	39.3% (38.1% - 40.4%)	49.5% (48.9% - 50.0%)
Female	49.2%	60.7% (59.6% - 61.9%)	50.5% (50.0% - 51.1%)
Race / Ethnicity			
White	73.0%	88.7% (87.4% - 89.0%)	75.4% (74.9% - 75.8%)
Black	7.7%	2.3% (1.9% - 2.6%)	7.0% (6.7% - 7.3%)
Hispanic	15.0%	5.3% (4.8% - 5.9%)	10.5% (10.1% - 10.8%)
Asian	1.6%	0.9% (0.7% - 1.2%)	0.5% (0.4% - 0.6%)
Other	2.7%	2.8% (2.4% - 3.2%)	6.6% (6.3% - 6.9%)

Prevalence Estimates Adjusted for Age and Gender

The county-wide prevalence estimates using emergency claims data was higher than the mailed survey for diabetes, but lower for asthma (Table 2). The correlation by sub-county area was very strong for these two conditions respectively at r = 0.90 (95% CI of 0.60 to 0.98) and r = 0.85 (95% CI of 0.44 to 0.97). For all other conditions except for diabetes, the county-wide prevalence estimates using emergency claims data was lower than the mailed survey. These correlations were graded across conditions: moderate for hypertension (r = 0.46, CI: -0.30 to 0.86) and COPD/emphysema (r = 0.42, CI: -0.34 to 0.85), and weak for cancer (r = 0.39, CI: -0.37 to 0.84) and hyperlipidemia (r = 0.23, CI: -0.51 to 0.78). Graphs of these correlations are found in Supplemental Figure 3, which demonstrate the variability between prevalence estimates especially for conditions with poor correlation by sub-county area. We displayed maps of prevalence estimates for diabetes, asthma, and hypertension based on survey results for the sub-county areas analyzed in Figure 1.

Table 2: Age and Gender Adjusted County-Level Disease Prevalence and Correlation at a Sub-County Level

Chronic Disease	County Wide Mailed Survey	Emergency Claims Data	Correlation Coefficient
Diabetes	12.7% (11.5% - 13.8%)	14.7% (14.3% - 15.1%)	0.90 (0.60 to 0.98)
Asthma	15.6% (14.1% - 17.1%)	8.7% (8.4% - 9.0%)	0.85 (0.44 to 0.97)
Hypertension	38.3%	36.1%	0.46

	(36.7% - 39.8%)	(35.6% - 36.6%)	(-0.30 to 0.86)
COPD / Emphysema	7.0% (6.2% - 7.9%)	4.0% (3.8% - 4.2%)	0.42 (-0.34 to 0.85)
Cancer	10.7% (9.8% - 11.5%)	3.9% (3.7% - 4.2%)	0.39 (-0.37 to 0.84)
Hyperlipidemia	33.9% (32.2% - 35.5%)	21.2% (20.8% - 21.7%)	0.23 (-0.51 to 0.78)

Emergency Department Surveillance

Among the 34,567 unique patients identified from emergency claims data, 76% had a geocodable home address, 20% were PO box only, and 4% were not geocodable but had a ZIP code located fully within Sullivan County. Using the 100 nearest neighbors among patients with a geocodable home address, we estimated unadjusted prevalence at the geocoded location of each patient and created interpolated rasters to provide a more geographically detailed maps of diabetes, asthma, and hypertension prevalence (Figure 2). These maps were able to identify localized clusters of disease throughout the county with greater geographic detail.

DISCUSSION

The intensity of health problems experienced by residents living in rural areas of the country underscores the need for improving our methods of health surveillance.[2 3] Our study's findings demonstrate a novel solution that uses emergency claims data to estimate chronic disease prevalence at a sub-county level. These estimates are important for identifying key hotspots of disease, which may reveal previously unexplored risk factors that increase disease burden in rural America and guide efforts to prevent chronic disease in specific geographic areas that experience the worst health outcomes.[26] Current health surveillance techniques rely on traditional methods such as telephone-based surveys. Not only are these methods costly and time-intensive, but due to dramatic shifts in phone use, response rates over the past two decades have dropped dramatically from around 36% to 9%.[27 28] The sample size of a large national health survey such as the BRFSS is inadequate for generating precise estimates of disease prevalence even at the county level for much of rural America, which is why the CDC has started to use alternative estimation methods to impute prevalence among rural counties.[29]

Recent efforts to provide greater geographic coverage have focused on approaches that use the data in adequately sampled areas and statistical models to extrapolate disease estimates for poorly sampled areas largely based on sociodemographic factors.[16] But many of these techniques have not been validated, and in the few instances when they have been compared, these approaches do not always work as well as expected.[6 7] Our mailed health survey found that countywide adjusted diabetes prevalence was 12.7%. This estimate is much higher than the CDC's most recent estimate of 9.5% in 2015, which is based on a modelling approach. For a given area, these modelling approaches can be especially imprecise when used to estimate disease prevalence in areas with low response rates, which includes many rural regions.

Other efforts to advance health surveillance methods have experimented with the use of claims data and electronic health records to provide estimates of disease prevalence. A recent study demonstrated that emergency claims data could be used to estimate chronic disease prevalence in New York City, and this approach was validated with results obtained from an annually performed citywide health survey. In this urban study, it was found that conditions including diabetes, hypertension, and asthma had correlations of 0.86, 0.88, and 0.77 respectively when analyzed among 34 sub-county areas.[14]

With our novel method of using emergency claims data to estimate chronic disease burden, we identified health records for a substantial majority of all adults in Sullivan County using five years of emergency claims data. Furthermore, the demographic patterns among this population of unique ED patients were much closer to Census estimates than our countywide mailed survey. Under-representation of certain demographic groups, especially minorities, is a common problem of traditional survey methods that can be adjusted for, as long as geographically matched sociodemographic data exists.[30] In rural areas where Census estimates for race and ethnicity often have wide confidence intervals, emergency claims data may provide an alternative population sample that closely mirrors the underlying population in a given region.

For some conditions such as diabetes and asthma, we found strong correlation between the two estimation methods for sub-county disease prevalence. For the other conditions studied, the strength of correlation was weaker. This may be attributable to disease-specific differences in the validity of both ED claims data and self-reported survey data. Prior research has shown that, for both data sources, validity is routinely higher for diabetes and asthma, but lower for other conditions like hyperlipidemia, with low sensitivity (i.e., under-diagnosis) being the reason for poor correlation.[8-10] It should be noted that though some conditions like COPD are a frequent primary diagnosis for a patient's ED visit, COPD may not be frequently accounted for as one of the secondary diagnoses, which are included in this ED-based surveillance approach.

Emergency claims data are already widely collected around the country, can capture a large population sample, and in some areas include address data that can be used to precisely identify where patients live. By geocoding these addresses, more precise health surveillance can provide detailed maps of disease burden. This granular level of geographic detail is important because localized hot spots of disease might otherwise be hidden as they are averaged out by neighboring areas of low disease prevalence. However, some important caveats should be understood before employing these methods. There is some variation in how accurately some hospitals capture chronic disease conditions using diagnosis codes.[31] In addition, for some parts of rural America, mail is only delivered to PO boxes, therefore the more geographically detailed maps of disease prevalence based on geocoded data may not be accurate in these regions where mail is not delivered directly. Furthermore, our study found substantial variability in prevalence estimates for conditions that may not be well captured by emergency claims data. More research may be needed to determine the best approaches for estimating disease prevalence in rural areas.

Limitations

Because surveys did not ask respondents to report household size, single-adult households are likely overrepresented. Furthermore, we did not specify a method of randomly selecting an adult in households with multiple residents, which may have contributed to bias in our sample. However, we age- and gender-standardized rates to the overall population in Sullivan County, which may partially reduce this bias. While our adjustment methods reduced age- and gender-specific non-response bias, we were unable to standardize by race and ethnicity due to the very small proportion of minorities in several areas of Sullivan County. Given that minorities often have higher rates of chronic disease and tend to have lower response rates, our mailed survey may have underestimated disease prevalence. Though the groups that frequently seek emergency care and those who respond to surveys tend to diverge, there may have been some sort of parallel bias that accounted for the correlations or disease prevalence identified in our study. Our method that used emergency claims data to estimate disease prevalence is subject to many of the limitations associated with the use of administrative data. Fidelity of coding some variables including race, ethnicity, and diagnosis codes can vary by hospital and may impact resulting disease prevalence estimates. Also, these claims data are often available about a year after they have been filed, thus there is some lag in reporting. In this study, emergency claims data were collected for 2011-2015, whereas the countywide survey was performed in 2017-2018.

CONCLUSION

We found that for select conditions, ED data may be useful for tracking disease prevalence in rural areas and may provide more geographically precise estimates. Given the infrastructure already in place to collect this data, efforts could be focused on collecting more accurate diagnosis codes and more detailed geographic data. This approach could potentially help match the limited health resources of a rural county to the geographic areas with the highest burden of disease.

Figure Legends

Figure 1. Sub-County Estimates of Adjusted Disease Prevalence Based on Mailed Survey Responses

Figure 2. Geographically Detailed Estimates of Unadjusted Disease Prevalence Based on Emergency Claims Data

Acknowledgments

The content of the study reflects the views of the authors and not the official position of the Sullivan County Public Health Department nor the NYU School of Medicine.

Author contributions

Study conception and design: DCL, LET, NAM; acquisition of the data: DCL, MO, MVN, AN, AJV; analysis and interpretation of the data: DCL, JMF, MO, CAK, CJS, AJV, LET; drafting of the manuscript: DCL; critical revision of the manuscript for intellectual content: JMF, MO, CAK, MVN, AN, CJS, AJV, LET. DCL is the guarantor of this work and had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This work was supported by the New York State Health Foundation grant number 16-04083. The funding source had no role in study design, conduct, data collection, data analysis, preparation of the manuscript, or the decision to submit the manuscript for publication.

Competing of interests

None declared.

Patient consent for publication

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination of our research.

Ethics approval

This study was approved by NYU School of Medicine's Institutional Review Board.

Data availability statement

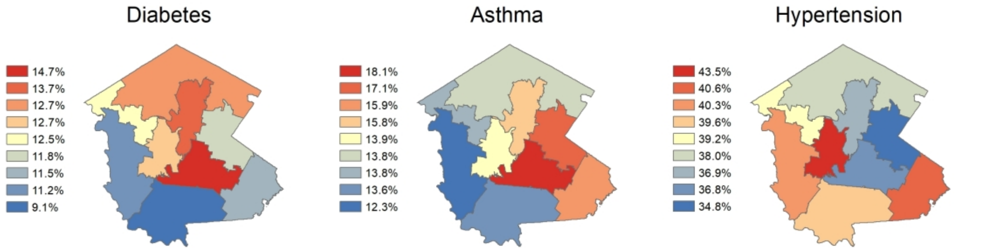
Deidentified participant data from the countywide health survey from 2017-2018 can be made available upon reasonable request by contacting Dr. David Lee at david.lee@nyumc.org so long as the requester agrees to the following conditions: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

REFERENCES

1. County Health Rankings and Roadmaps. 2018 County Health Rankings Report Madison, WI: University of Wisconsin Population Health Institute; 2018 [Available from: http://www.countyhealthrankings.org/sites/default/files/state/downloads/CHR2018_NY.pdf accessed November 8th 2018.
2. Harris JK, Beatty K, Leider JP, et al. The Double Disparity Facing Rural Local Health Departments. *Annual review of public health* 2016;37:167-84. doi: 10.1146/annurev-publhealth-031914-122755 [published Online First: 2016/01/07]
3. Brundisini F, Giacomini M, DeJean D, et al. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. *Ontario health technology assessment series* 2013;13(15):1-33. [published Online First: 2013/11/15]
4. Puma JE, Belansky ES, Garcia R, et al. A Community-Engaged Approach to Collecting Rural Health Surveillance Data. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association* 2017;33(3):257-65. doi: 10.1111/jrh.12185 [published Online First: 2016/06/16]
5. Bellamy GR, Bolin JN, Gamm LD. Rural Healthy People 2010, 2020, and beyond: the need goes on. *Family & community health* 2011;34(2):182-8. doi: 10.1097/FCH.0b013e31820dea1c [published Online First: 2011/03/08]
6. Feldman JM. Re: Obesity and place: Chronic disease in the 500 largest U.S. cities. *Obesity research & clinical practice* 2018;12(5):490. doi: 10.1016/j.orcp.2018.08.003 [published Online First: 2018/09/18]
7. Wang Y, Holt JB, Zhang X, et al. Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013. *Preventing chronic disease* 2017;14:E99. doi: 10.5888/pcd14.170281 [published Online First: 2017/10/20]
8. Hebert PL, Geiss LS, Tierney EF, et al. Identifying persons with diabetes using Medicare claims data. *American journal of medical quality : the official journal of the American College of Medical Quality* 1999;14(6):270-7. doi: 10.1177/106286069901400607 [published Online First: 2000/01/07]
9. Tessier-Sherman B, Galusha D, Taiwo OA, et al. Further validation that claims data are a useful tool for epidemiologic research on hypertension. *BMC public health* 2013;13:51. doi: 10.1186/1471-2458-13-51 [published Online First: 2013/01/22]
10. Robinson JR, Young TK, Roos LL, et al. Estimating the burden of disease. Comparing administrative data and self-reports. *Medical care* 1997;35(9):932-47. [published Online First: 1997/09/23]
11. Colantonio LD, Levitan EB, Yun H, et al. Use of Medicare Claims Data for the Identification of Myocardial Infarction: The Reasons for Geographic And Racial Differences in Stroke (REGARDS) Study. *Medical care* 2018 doi: 10.1097/mlr.0000000000001004 [published Online First: 2018/10/27]
12. Hope KG, Merritt TD, Durrheim DN, et al. Evaluating the utility of emergency department syndromic surveillance for a regional public health service. *Communicable diseases intelligence quarterly report* 2010;34(3):310-8. [published Online First: 2010/11/26]

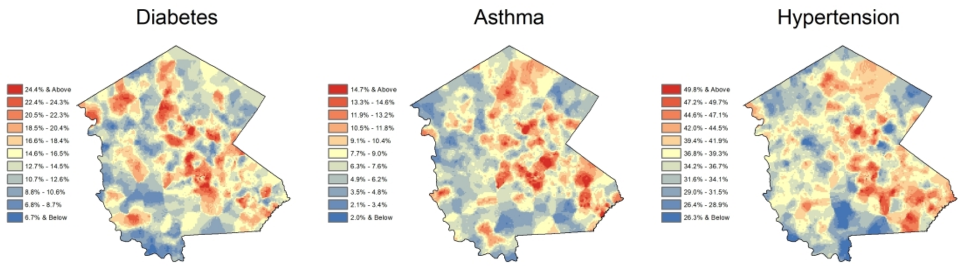
13. Centers for Disease Control and Prevention. Health, United States, 2013: With Special Feature on Prescription Drugs. Hyattsville, MD: National Center for Health Statistics, 2014.
14. Lee DC, Long JA, Wall SP, et al. Determining Chronic Disease Prevalence in Local Populations Using Emergency Department Surveillance. *American journal of public health* 2015;105(9):e67-74. doi: 10.2105/ajph.2015.302679 [published Online First: 2015/07/17]
15. Centers for Disease Control and Prevention. Public Health Surveillance in Rural Communities 2019 [Available from: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/surveillance-in-rural-communities.htm.
16. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. *American journal of preventive medicine* 2004;26(5):453-60. doi: 10.1016/j.amepre.2004.02.004 [published Online First: 2004/05/29]
17. Zhang X, Holt JB, Yun S, et al. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American journal of epidemiology* 2015;182(2):127-37. doi: 10.1093/aje/kwv002 [published Online First: 2015/05/10]
18. Coburn AF, MacKinney AC, McBride TD, et al. Choosing Rural Definitions: Implications for Health Policy Iowa City, IA: Rural Policy Research Institute Health Panel, 2007.
19. Centers for Disease Control and Prevention. SMART BRFSS MMSA Methodology Atlanta, GA2016 [Available from: https://www.cdc.gov/brfss/smart/2016/2016_SMART_BRFSS_MMSA_Methodology.pdf accessed November 8th 2018.
20. NYS ITS GIS Program Office. NYS GIS Clearinghouse Albany, NY [Available from: www.gis.ny.gov accessed November 8th 2018.
21. Mullen PD, Savas LS, Bundy LT, et al. Minimal intervention delivered by 2-1-1 information and referral specialists promotes smoke-free homes among 2-1-1 callers: a Texas generalisation trial. *Tobacco control* 2016;25(Suppl 1):i10-i18. doi: 10.1136/tobaccocontrol-2016-053045 [published Online First: 2016/10/05]
22. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. *BMC medical research methodology* 2013;13:49. doi: 10.1186/1471-2288-13-49 [published Online First: 2013/03/26]
23. Lee DC, Doran KM, Polsky D, et al. Geographic variation in the demand for emergency care: A local population-level analysis. *Healthcare (Amsterdam, Netherlands)* 2016;4(2):98-103. doi: 10.1016/j.hjdsi.2015.05.003 [published Online First: 2016/06/28]
24. Klein RJ, Schoenborn CA. Age adjustment using the 2000 projected U.S. population. Healthy People Statistical Notes. Hyattsville, Maryland: National Center for Health Statistics, January 2001.
25. American Association for Public Opinion Research. Mailed Surveys of Unnamed Persons Oakbrook Terrace, ILApril 2016 [Available from:

https://www.aapor.org/getattachment/Standards-Ethics/Standard-Definitions-(1)/Mail_Surveys_Unnamed_Persons_042216.pdf.aspx accessed November 8th 2018.
26. Brownson RC, Fielding JE, Green LW. Building Capacity for Evidence-Based Public Health: Reconciling the Pulls of Practice and the Push of Research. *Annual review of public health* 2018;39:27-53. doi: 10.1146/annurev-publhealth-040617-014746 [published Online First: 2017/11/23]
27. Galea S, Tracy M. Participation rates in epidemiologic studies. *Annals of epidemiology* 2007;17(9):643-53. doi: 10.1016/j.annepidem.2007.03.013 [published Online First: 2007/06/08]
28. Keeter S, Hatley N, Kennedy C, et al. What Low Response Rates Mean for Telephone Surveys Washington DC: Pew Research Center; 2017 [Available from: http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/ accessed July 25th 2017.
29. O'Connor A, Wellenius G. Rural-urban disparities in the prevalence of diabetes and coronary heart disease. *Public health* 2012;126(10):813-20. doi: 10.1016/j.puhe.2012.05.029 [published Online First: 2012/08/28]
30. Link MW, Mokdad AH, Stackhouse HF, et al. Race, ethnicity, and linguistic isolation as determinants of participation in public health surveillance surveys. *Preventing chronic disease* 2006;3(1):A09. [published Online First: 2005/12/17]
31. Lee DC, Swartz JL, Koziatek CA, et al. Assessing the Reliability of Performing Citywide Chronic Disease Surveillance Using Emergency Department Data from Sentinel Hospitals. *Population health management* 2017 doi: 10.1089/pop.2016.0168 [published Online First: 2017/03/25]



Sub-County Estimates of Adjusted Disease Prevalence Based on Mailed Survey Responses

228x76mm (300 x 300 DPI)



Geographically Detailed Estimates of Unadjusted Disease Prevalence Based on Emergency Claims Data

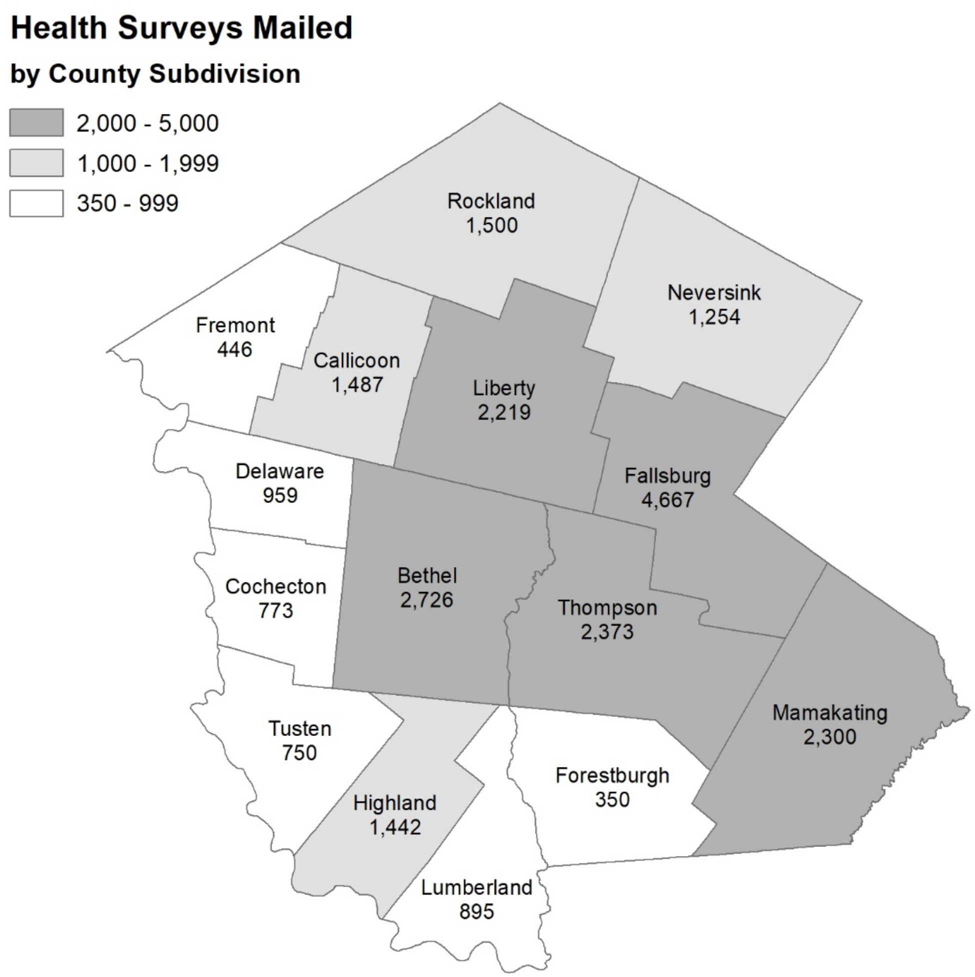
304x101mm (300 x 300 DPI)

Supplemental Table 1: Households, Mailings, and Responses by Geographic Level in Sullivan County

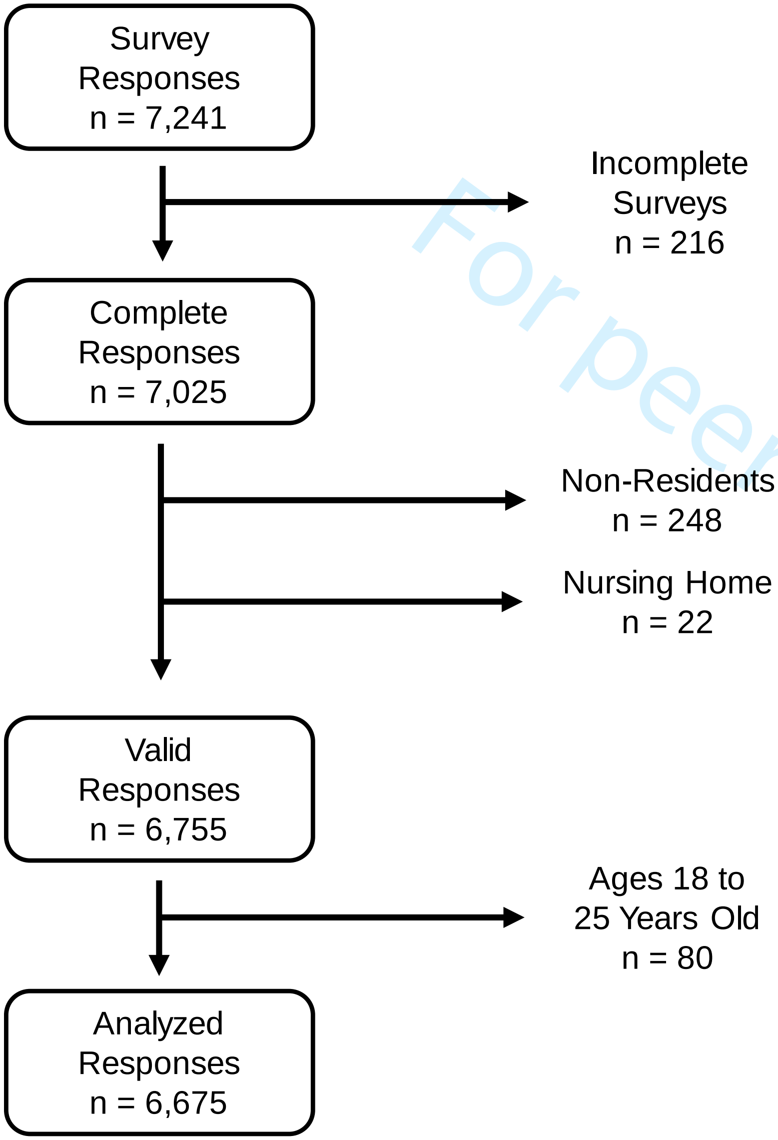
Zip Code	Population	Households	Mailings	Responses	Subdivison	Population	Households	Mailings	Responses	Subcounty	Population	Households	Mailings	Responses
12720	143	159	159	51	Bethel	2,900	2,774	2,726	719	Bethel	2,900	2,774	2,726	719
12749	266	382	382	97	Bethel					Bethel				
12762	288	323	323	104	Bethel					Bethel				
12778	437	618	618	179	Bethel					Bethel				
12783	1,396	798	750	157	Bethel					Bethel				
12786	370	494	494	131	Bethel					Bethel				
12724	217	199	199	85	Callicoon	2,175	1,731	1,487	526	Callicoon / Fremont	2,817	2,177	1,933	707
12748	1,172	994	750	205	Callicoon					Callicoon / Fremont				
12766	309	179	179	68	Callicoon					Callicoon / Fremont				
12791	477	359	359	168	Callicoon					Callicoon / Fremont				
12736	89	84	84	27	Fremont	642	446	446	181	Callicoon / Fremont				
12741	157	159	159	69	Fremont					Callicoon / Fremont				
12760	353	152	152	56	Fremont					Callicoon / Fremont				
12767	43	51	51	29	Fremont					Callicoon / Fremont				
12726	959	551	551	186	Cochecton	1,142	773	773	248	Cochecton / Delaware / Tusten	4,139	2,842	2,482	763
12752	183	222	222	62	Cochecton					Cochecton / Delaware / Tusten				
12723	1,391	893	750	212	Delaware	1,711	1,102	959	289	Cochecton / Delaware / Tusten				
12745	127	92	92	36	Delaware					Cochecton / Delaware / Tusten				
12750	193	117	117	41	Delaware					Cochecton / Delaware / Tusten				
12764	1,286	967	750	226	Tusten	1,286	967	750	226	Cochecton / Delaware / Tusten				
12733	1,070	550	550	79	Fallsburg	8,849	5,737	4,667	966	Fallsburg	8,849	5,737	4,667	966
12738	327	132	132	34	Fallsburg					Fallsburg				
12747	1,143	735	735	170	Fallsburg					Fallsburg				
12759	886	574	574	132	Fallsburg					Fallsburg				
12763	409	426	426	129	Fallsburg					Fallsburg				
12779	1,244	1,162	750	114	Fallsburg					Fallsburg				
12788	2,381	1,005	750	180	Fallsburg					Fallsburg				
12789	1,389	1,153	750	128	Fallsburg					Fallsburg				
12729	82	41	41	16	Forestburgh	781	350	350	122	Forestburgh /. Highland / Lumberland	4,314	2,763	2,687	856
12777	699	309	309	106	Forestburgh					Forestburgh /. Highland / Lumberland				
12719	1,042	520	520	179	Highland	2,013	1,442	1,442	493	Forestburgh /. Highland / Lumberland				
12732	577	480	480	155	Highland					Forestburgh /. Highland / Lumberland				
12743	242	184	184	66	Highland					Forestburgh /. Highland / Lumberland				
12792	152	258	258	93	Highland					Forestburgh /. Highland / Lumberland				
12737	1,382	826	750	193	Lumberland	1,520	971	895	241	Forestburgh /. Highland / Lumberland				
12770	138	145	145	48	Lumberland					Forestburgh /. Highland / Lumberland				
12734	697	652	652	162	Liberty	6,119	4,875	2,219	620	Liberty	6,119	4,875	2,219	620
12754	4,702	3,406	750	178	Liberty					Liberty				
12768	532	556	556	175	Liberty					Liberty				
12787	188	261	261	105	Liberty					Liberty				
12721	4,153	2,108	750	166	Mamakating	8,519	5,368	2,300	578	Mamakating	8,519	5,368	2,300	578
12722	77	200	200	59	Mamakating					Mamakating				
12769	156	100	100	40	Mamakating					Mamakating				
12781	211	133	133	42	Mamakating					Mamakating				
12785	704	367	367	63	Mamakating					Mamakating				
12790	3,218	2,460	750	208	Mamakating					Mamakating				
12725	143	103	103	24	Neversink	2,046	1,286	1,254	400	Neversink / Rockland	6,736	4,134	2,754	789
12740	1,208	782	750	236	Neversink					Neversink / Rockland				
12765	695	401	401	140	Neversink					Neversink / Rockland				
12758	2,966	1,997	750	204	Rockland	4,690	2,848	1,500	389	Neversink / Rockland				
12776	1,724	851	750	185	Rockland					Neversink / Rockland				
12701	7,795	6,044	750	228	Thompson	10,104	8,414	2,373	677	Thompson	10,104	8,414	2,373	677
12742	62	197	197	66	Thompson					Thompson				
12751	561	564	564	174	Thompson					Thompson				
12775	1,648	1,497	750	170	Thompson					Thompson				
12784	38	112	112	39	Thompson					Thompson				
Total	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**

Note: Population estimates are based on ZCTAs and may not align with ZIP codes. A population adjustment was required in ZIP code 12729 due to gross mismatch between ACS estimate and number of households from sampling frame.

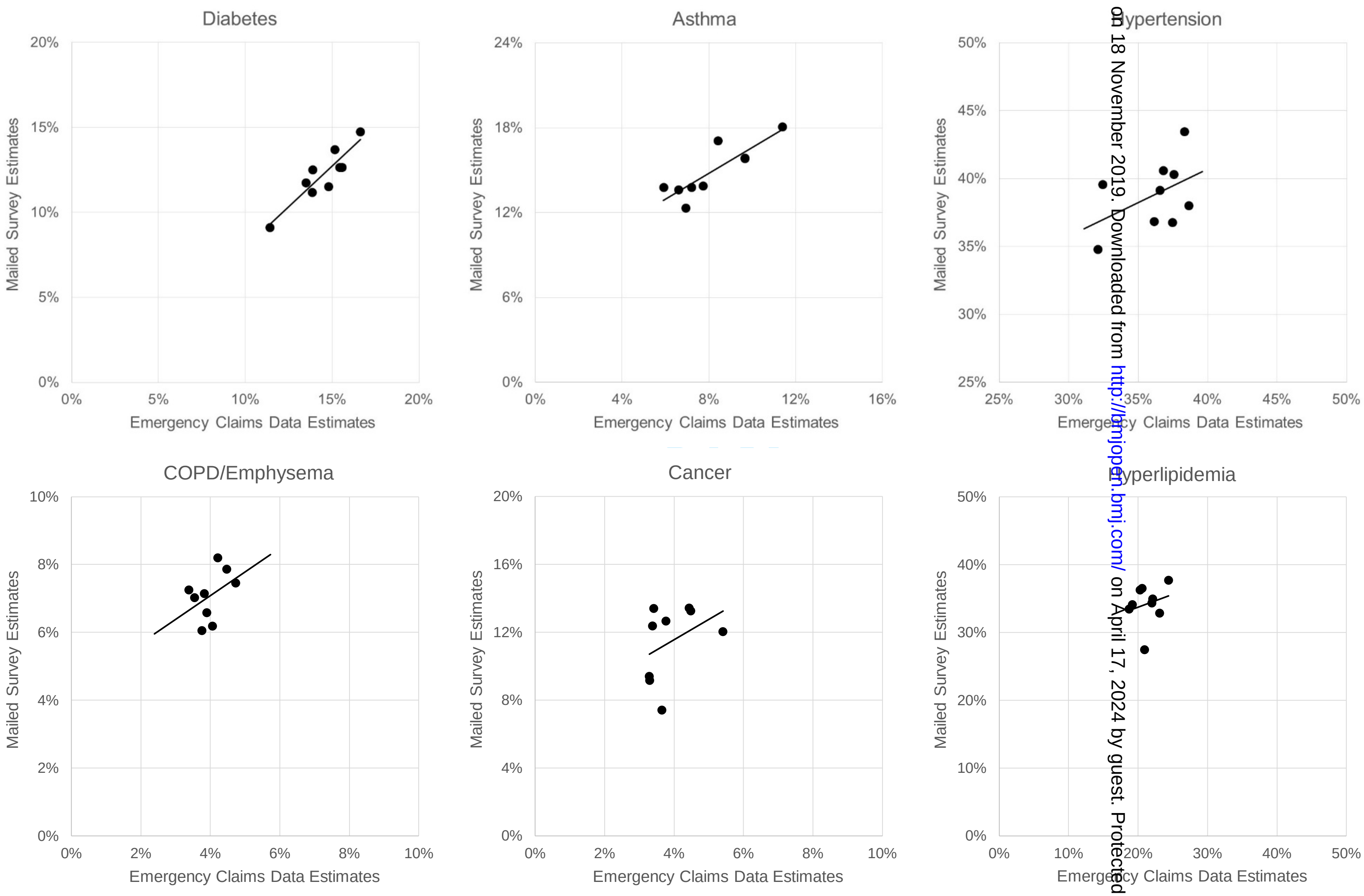
Supplemental Figure 1: Health Survey Mailed by County Subdivision



Supplemental Figure 2: Diagram of Survey Responses and Exclusions



Supplemental Figure 3: Comparison of Sub-County Estimates of Disease Prevalence



STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **Page 1**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **Pages 2-3**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **Pages 4-5**
Objectives	3	State specific objectives, including any prespecified hypotheses **Page 5, Paragraph 2**
Methods		
Study design	4	Present key elements of study design early in the paper **Page 6, Paragraph 1**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **Pages 6-7**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants **Pages 6-7**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **Page 8, Paragraph 1**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **Pages 8-9**
Bias	9	Describe any efforts to address potential sources of bias **Pages 8-9**
Study size	10	Explain how the study size was arrived at **Page 11**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **Pages 8-9**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **Pages 8-10**
		(*b*) Describe any methods used to examine subgroups and interactions **N/A**
		(*c*) Explain how missing data were addressed **Page 11**
		(*d*) If applicable, describe analytical methods taking account of sampling strategy **Pages 8-9**
		(*e*) Describe any sensitivity analyses **N/A**
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **Page 11**
		(b) Give reasons for non-participation at each stage **Page 11**
		(c) Consider use of a flow diagram **Supplemental Materials**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **Page 11, Paragraph 2**
		(b) Indicate number of participants with missing data for each variable of interest **Page 11, Supplemental Materials**
Outcome data	15*	Report numbers of outcome events or summary measures **Page 13**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and

		their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **Page 13**
		(*b*) Report category boundaries when continuous variables were categorized **N/A**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period **N/A**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **Page 14**
Discussion		
Key results	18	Summarise key results with reference to study objectives **Page 15, Paragraph 1**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **Page 18, Paragraph 1**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **Pages 15-17**
Generalisability	21	Discuss the generalisability (external validity) of the study results **Page 18**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **Page 20**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data: A Cross-Sectional Comparison of Survey versus Claims Data in Sullivan County, New York

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-033373.R2
Article Type:	Original research
Date Submitted by the Author:	31-Oct-2019
Complete List of Authors:	Lee, David; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine; New York University School of Medicine, Department of Population Health Feldman, Justin; New York University School of Medicine, Department of Population Health Osorio, Marcela; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Koziatek, Christian; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nguyen, Michael; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Nagappan, Ashwini; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Shim, Christopher; California Northstate University College of Medicine Vinson, Andrew; New York University School of Medicine, Ronald O. Perelman Department of Emergency Medicine Thorpe, Lorna; New York University School of Medicine, Department of Population Health McGraw, Nancy; Sullivan County Public Health Services
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Diabetes and endocrinology, Emergency medicine, Epidemiology, Public health, Research methods
Keywords:	EPIDEMIOLOGY, PUBLIC HEALTH, HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™ Manuscripts

Title: Improving the Geographic Precision of Rural Chronic Disease Surveillance by Using Emergency Claims Data: A Cross-Sectional Comparison of Survey versus Claims Data in Sullivan County, New York

Authors

David C. Lee, MD, MS[1,2]
Justin M. Feldman, MPH, ScD[2]
Marcela Osorio[1]
Christian A. Koziatek, MD[1]
Michael V. Nguyen[1]
Ashwini Nagappan[1]
Christopher J. Shim, JD, MS[3]
Andrew J. Vinson, MA, GISP[1]
Lorna E. Thorpe, PhD, MPH[2]
Nancy A. McGraw, LCSW, MBA, MPH[4]

Author Affiliations

1. Ronald O. Perelman Department of Emergency Medicine, New York University School of Medicine, New York, NY, USA
2. Department of Population Health, New York University School of Medicine, New York, NY, USA
3. California Northstate University College of Medicine, Elk Grove, CA, USA
4. Sullivan County Public Health Department, Liberty, NY, USA

Corresponding Author

David C. Lee, MD, MS
227 East 30th Street, Room 107
New York, NY 10016
Phone: (212) 562-6561
Email: David.Lee@nyulangone.org

Manuscript Word count: 3,380

ABSTRACT

Objectives: Some of the most pressing health problems are found in rural America. However, the surveillance needed to track and prevent disease in these regions is lacking. Our objective was to perform a comprehensive health survey of a single rural county to assess the validity of using emergency claims data to estimate rural disease prevalence at a sub-county level.

Design: We performed a cross-sectional study of chronic disease prevalence estimates using emergency department claims data versus mailed health surveys designed to capture a substantial proportion of residents in New York's rural Sullivan County.

Setting: Sullivan County, a rural county ranked second-to-last for health outcomes in New York State.

Participants: Adult residents of Sullivan County aged 25 years and older who responded to the health survey in 2017-2018 or had at least one emergency department visit in 2011-2015.

Outcome Measures: We compared age- and gender-adjusted prevalence of hypertension, hyperlipidemia, diabetes, cancer, asthma, and COPD/emphysema among nine sub-county areas.

Results: Our countywide mailed survey obtained 6,675 completed responses for a response rate of 30.4%. This sample represented more than 12% of the estimated 53,020 adults in Sullivan County. Using emergency claims data, we identified 34,576 adults from Sullivan County who visited an emergency department at least once during 2011-2015. At a sub-county level, prevalence estimates from mailed surveys and emergency claims data correlated especially well for diabetes (r=0.90) and asthma (r=0.85). Other conditions were not as well correlated (range: 0.23-0.46). Using emergency claims data, we created more geographically detailed maps of disease prevalence using geocoded addresses.

Conclusions: For select conditions, emergency claims data may be useful for tracking disease prevalence in rural areas and provide more geographically detailed estimates. For rural regions lacking robust health surveillance, emergency claims data can inform how to geographically target efforts to prevent chronic disease.

Article Summary

Strengths and limitations of this study

- Validates the use of emergency claims data to perform geographically detailed surveillance in rural settings.
- Provides a standard for estimating disease prevalence at a local level by performing a countywide mailed survey.
- Limited by the accuracy of diagnosis codes found in claims data and is more accurate for conditions likely to be captured during emergency visits.
- Has the potential to improve rural health surveillance by using existing data to track the burden of chronic diseases.

INTRODUCTION

In New York State, Sullivan County has been ranked 61st out of 62 by the County Health Rankings based on rates of premature death and quality of life (poor overall, physical, mental health and low birthweights) just behind Bronx County in New York City.[1] Located just two hours northwest of New York City, Sullivan County is rural and more than 70% of its residents are White. Like many rural areas of America, Sullivan County has faced significant economic challenges, along with disparities in healthcare access.[2 3] Though some of the most pressing health problems can be found in rural America, their public health institutions lack timely data needed to provide geographically detailed chronic disease surveillance.[4 5] Nationwide health surveys, such as the Behavioral Risk Factor Surveillance System (BRFSS), often have inadequate coverage of these rural regions, and efforts to use models to extrapolate estimates of disease prevalence have questionable validity.[6 7]

In recent years, there has been increasing interest in using alternative sources of data to track chronic disease prevalence.[8-10] Approaches using claims data and electronic health records have emerged among the potential options.[11] These data are collected routinely by state agencies and may provide a cost-effective, ready-to-analyze alternative to expensive and time-intensive traditional survey methods.[12] For instance, 1 in 5 Americans report having visited an emergency department (ED) in the past year, which provides a 20% population sample with a single year of data.[13 14] However, these approaches need to be validated before widespread dissemination because, unlike surveys, they are not random population samples and may therefore not be representative.

There are several challenges that make estimation of chronic disease prevalence in rural areas difficult. The Centers for Disease Control and Prevention (CDC) has started to use modelling approaches with Bayesian and spatial smoothing with BRFSS data to estimate county-level disease prevalence in rural areas.[15] But, few traditional health surveys have been performed in these areas with sample sizes adequate for sub-county level area estimation.[16 17] In addition, there is similarly sparse data on the sociodemographic composition of these rural regions (as evidenced by wide confidence intervals for sub-county estimates of race and ethnicity). Furthermore, ZIP codes, county borders, and other geographic units are less likely to align in rural areas, limiting the possibility of attributing aggregated data to specific regions.[18] In addition, certain traditional survey techniques used to refine estimates based on underlying demographic characteristics (e.g., statistical weighting, adjustment, or stratification) that are often performed with Census data cannot be readily applied in rural areas especially if there is insufficient data.[19]

The goal of this study was to perform a comprehensive health survey of a single rural county in the United States. We report the results of a geographically distributed health survey delivered by mail to households within Sullivan County, New York. We then compare the disease prevalence estimates obtained from these surveys to a novel method that uses emergency claims data to identify areas with a higher burden of chronic disease.[14]

METHODS

Study Design

We administered a brief health survey by mail throughout Sullivan County during Fall 2017 and Spring 2018 to a random sample of residential addresses. We used survey data to estimate the age- and gender-standardized prevalence of several chronic diseases at a sub-county level. We also estimated disease prevalence using the comprehensive, all-payer New York Statewide Planning and Research Cooperative System (SPARCS) claims database. Our alternative measure was the proportion of ED patients with ≥ 1 diagnostic code for a given disease on ≥ 1 emergency visit during the period 2011-2015.[14] In each method, residents with an address located at a nursing or correctional facility were excluded to estimate prevalence for the non-institutionalized population. This study was approved by NYU School of Medicine's Institutional Review Board.

Mailed Health Surveys

To generate a sampling frame for our mailed health survey, we obtained point and parcel data for all mailing addresses in Sullivan County from the New York State GIS Clearinghouse (www.gis.ny.gov).[20] This data source was selected because it contained property class and land use data. Addresses were filtered to include any residential listing not marked as seasonal housing. We also included commercial addresses listed as apartments. This list of mailing addresses was then refined using an address verification service (www.smartystreets.com) to select valid, non-vacant mailable addresses.[21] As a substantial proportion of residents do not receive delivered mail in Sullivan County, we also queried the address verification service to find all valid, mailable PO boxes in the county. The final sampling frame consisted of 39,084 households located across 56 ZIP codes within Sullivan County.

Given the sparse population in some areas of the rural county, less-populated ZIP codes were oversampled to maximize the geographic coverage of the survey over the entire county. To do so, we used a quota sampling strategy. We mailed surveys to a random sample of 750 households for each ZIP code in our sampling frame. In ZIP codes with fewer than 750 households, all households were mailed a survey. Each health survey consisted of questions that first confirmed residence within Sullivan County and age over 18 years, and then asked a brief selection of health and demographic questions derived from the BRFSS (See Supplemental File 1 for mailed health survey).[22] For households with multiple residents, we asked that only one adult respond to the survey. We mailed a survey to 24,141 or 62% of the households in our sampling frame. Survey respondents were offered a $10 gift card for participation and a stamped return envelope was enclosed in the surveys. The Sullivan County Public Health Department also made local news outlets aware of the survey and fielded phone calls from local residents to confirm that the survey was legitimate.

Emergency Claims Data

Using the SPARCS, all-payer claims database, we identified all adult patients who had visited an ED located at a general acute care hospital in New York State between 2011 and 2015. We included all patients with a PO box or home address located within the borders of Sullivan County. Patients with more than one ED visit either at the same hospital or different hospitals were counted as a single observation by collapsing multiple visits using unique identifiers from SPARCS. The result was a listing of unique Sullivan County residents who had accessed emergency care at least once during the five-year period.

Study Outcomes

Our primary outcome was the prevalence of chronic disease at a sub-county level as identified by our mailed health survey or estimated using emergency claims data. In our mailed survey, respondents were asked if they had ever been diagnosed with hypertension, hyperlipidemia, cancer, diabetes, asthma, chronic obstructive pulmonary disease (COPD) or emphysema. In our analysis of emergency claims data, all available primary and secondary diagnosis codes across visits were scanned by individual for the presence of ≥ 1 diagnosis code during ≥ 1 ED visit for these same conditions. The codes from the International Classification of Diseases (ICD-9 and ICD-10) used were: hypertension (401-405 or I10-I16), hyperlipidemia (272 or E78), diabetes (250 or E10-E11), cancer (140-239 or C00-C96), asthma (493 or J45), and COPD/emphysema (491-492 or J43-J44). Thus, prevalence was estimated as a proportion: the number of unique ED patients with each of the listed conditions, divided by the total number of unique ED patients.

Statistical Analysis

To generate the sub-county areas in our analysis, we first grouped ZIP codes based on the Census-defined subdivisions (i.e., town borders) within Sullivan County. ZIP codes were assigned to these subdivisions based on the largest area of overlap given that ZIP code boundaries do not exactly align with town borders.[18] After grouping ZIP codes into these fifteen subdivisions, it was found that ten of these subdivisions had less than 2,000 households that received a mailed survey and were thus unlikely to obtain the minimum 500 survey responses, a benchmark set by the CDC for obtaining acceptably narrow confidence intervals for prevalence estimation (Supplemental Figure 1).[19] Therefore, these less populated subdivisions were systematically merged with each other based on proximity and population density to form four sub-county areas with a sufficient number of sampled households. The result was nine sub-county areas made of five subdivisions with adequate sampling and four areas combining neighboring subdivisions to attain adequate sampling. (See Supplemental Table 1 for more details of aggregating ZIP codes into subdivisions and then sub-county areas).

In aggregating prevalence estimates between ZIP codes to create the sub-county areas, we used two weighting approaches. For the mailed survey, we applied design weights (the inverse probability of selection from the sampling frame) to account for our oversampling of less-populated ZIP codes. For the emergency claims data, we weighted ZIP code prevalence estimates by the inverse of the total number of unique ED patients divided by the Census estimate of adults aged 25 years and older for each ZIP code in Sullivan County to account for known differences in ED utilization based on proximity to the nearest hospital.[23] Prevalence estimates using both methods were then standardized to the overall age and gender distributions in Sullivan County from the most recent five-year 2012-2016 American Community Survey (ACS).[24] We then calculated Pearson correlation coefficients comparing the prevalence estimates obtained using the two methods at the sub-county level. By convention, the strength of correlation was graded as very strong (0.80 - 1.00), strong (0.60 - 0.79), moderate (0.40 - 0.59), weak (0.20 - 0.39), and very weak (0.00 - 0.19).

Geographic Analysis

We also performed geographically detailed surveillance using the larger sample of Sullivan County residents identified in emergency claims data. For the subset of patients with a geocodable home address, we calculated unadjusted disease prevalence among their 100 nearest neighbors identified in the population of unique ED patients. We then interpolated rasters from this point data using the inverse squared distance technique. Chronic disease prevalence maps were generated from these unadjusted prevalence estimations for diabetes, asthma, and hypertension, with categories based on standard deviations from the mean. For comparison, these maps were also created based on the 200 nearest neighbors to assess the influence of changing this parameter.

Statistical analyses were performed using Stata 14.2 (Statacorp; College Station, TX, 2015). Geographic analysis and mapping were performed using ArcGIS Desktop 10.5.1 (ESRI; Redlands, CA, 2017).

RESULTS

Mailed Survey Responses

Of the 24,141 surveys that we mailed to addresses within Sullivan County, approximately 20% were returned to sender even after using an address verification service (Supplemental Figure 2). Of the 7,241 survey responses received, 216 were missing key demographic information or were otherwise incomplete, 248 were not residents of Sullivan County, and 22 were located at a nursing or correctional facility. In addition, only 80 respondents were aged 18 to 24 years old, which we deemed too few for inclusion in the study. Therefore, we limited study results to adults aged 25 years and older. Using the AAPOR RR2 definition for mail surveys of unnamed persons, our response rate was 30.4%.[25]

Population Characteristics

The countywide mailed survey received valid responses for 6,675 adults or 12.6% of the adult population 25 years and older in Sullivan County. Using five years of emergency claims data, we were able to identify 65.2% of the Census-estimated adult population 25 years and older in Sullivan County. In comparison to ACS 2012-2016 Census estimates, survey respondents were notably older (42.5% versus Census estimate of 23.9% aged 65 years and older). In comparison, the population of unique ED patients was slightly younger (39.2% versus Census estimate of 33.0% aged 25 to 44 years old). A higher proportion of survey respondents were female (60.7% versus Census estimate of 49.2%). Also, a higher proportion of survey respondents were non-Hispanic white (88.7% versus Census estimate of 73.0%). However, the sex and race/ethnicity distributions of the unique ED patient population were similar to Census estimates (Table 1).

Table 1: Demographic Comparisons Among Census Estimates and Data Sources

Demographic Comparisons	2012-2016 Census Estimates	County Wide Mailed Survey	Emergency Claims Data
Total	53,020	6,675	34,567
Age			
25 to 44	33.0%	15.9% (15.0% - 16.8%)	39.2% (38.7% - 39.7%)
45 to 64	43.1%	41.6% (40.4% - 42.7%)	36.9% (36.4% - 37.5%)
65 and older	23.9%	42.5% (41.3% - 43.7%)	23.9% (23.4% - 24.3%)
Sex			
Male	50.8%	39.3% (38.1% - 40.4%)	49.5% (48.9% - 50.0%)
Female	49.2%	60.7% (59.6% - 61.9%)	50.5% (50.0% - 51.1%)
Race / Ethnicity			
White	73.0%	88.7% (87.4% - 89.0%)	75.4% (74.9% - 75.8%)
Black	7.7%	2.3% (1.9% - 2.6%)	7.0% (6.7% - 7.3%)
Hispanic	15.0%	5.3% (4.8% - 5.9%)	10.5% (10.1% - 10.8%)
Asian	1.6%	0.9% (0.7% - 1.2%)	0.5% (0.4% - 0.6%)
Other	2.7%	2.8% (2.4% - 3.2%)	6.6% (6.3% - 6.9%)

Prevalence Estimates Adjusted for Age and Gender

The county-wide prevalence estimates using emergency claims data was higher than the mailed survey for diabetes, but lower for asthma (Table 2). The correlation by sub-county area was very strong for these two conditions respectively at r = 0.90 (95% CI of 0.60 to 0.98) and r = 0.85 (95% CI of 0.44 to 0.97). For all other conditions except for diabetes, the county-wide prevalence estimates using emergency claims data was lower than the mailed survey. These correlations were graded across conditions: moderate for hypertension (r = 0.46, CI: -0.30 to 0.86) and COPD/emphysema (r = 0.42, CI: -0.34 to 0.85), and weak for cancer (r = 0.39, CI: -0.37 to 0.84) and hyperlipidemia (r = 0.23, CI: -0.51 to 0.78). Graphs of these correlations are found in Supplemental Figure 3, which demonstrate the variability between prevalence estimates especially for conditions with poor correlation by sub-county area. We displayed maps of prevalence estimates for diabetes, asthma, and hypertension based on survey results for the sub-county areas analyzed in Figure 1.

Table 2: Age and Gender Adjusted County-Level Disease Prevalence and Correlation at a Sub-County Level

Chronic Disease	County Wide Mailed Survey	Emergency Claims Data	Correlation Coefficient
Diabetes	12.7% (11.5% - 13.8%)	14.7% (14.3% - 15.1%)	0.90 (0.60 to 0.98)
Asthma	15.6% (14.1% - 17.1%)	8.7% (8.4% - 9.0%)	0.85 (0.44 to 0.97)
Hypertension	38.3%	36.1%	0.46

	(36.7% - 39.8%)	(35.6% - 36.6%)	(-0.30 to 0.86)
COPD / Emphysema	7.0% (6.2% - 7.9%)	4.0% (3.8% - 4.2%)	0.42 (-0.34 to 0.85)
Cancer	10.7% (9.8% - 11.5%)	3.9% (3.7% - 4.2%)	0.39 (-0.37 to 0.84)
Hyperlipidemia	33.9% (32.2% - 35.5%)	21.2% (20.8% - 21.7%)	0.23 (-0.51 to 0.78)

Emergency Department Surveillance

Among the 34,567 unique patients identified from emergency claims data, 76% had a geocodable home address, 20% were PO box only, and 4% were not geocodable but had a ZIP code located fully within Sullivan County. Using the 100 nearest neighbors among patients with a geocodable home address, we estimated unadjusted prevalence at the geocoded location of each patient and created interpolated rasters to provide a more geographically detailed maps of diabetes, asthma, and hypertension prevalence (Figure 2). These maps were able to identify localized clusters of disease throughout the county with greater geographic detail.

DISCUSSION

The intensity of health problems experienced by residents living in rural areas of the country underscores the need for improving our methods of health surveillance.[2 3] Our study's findings demonstrate a novel solution that uses emergency claims data to estimate chronic disease prevalence at a sub-county level. These estimates are important for identifying key hotspots of disease, which may reveal previously unexplored risk factors that increase disease burden in rural America and guide efforts to prevent chronic disease in specific geographic areas that experience the worst health outcomes.[26] Current health surveillance techniques rely on traditional methods such as telephone-based surveys. Not only are these methods costly and time-intensive, but due to dramatic shifts in phone use, response rates over the past two decades have dropped dramatically from around 36% to 9%.[27 28] The sample size of a large national health survey such as the BRFSS is inadequate for generating precise estimates of disease prevalence even at the county level for much of rural America, which is why the CDC has started to use alternative estimation methods to impute prevalence among rural counties.[29]

Recent efforts to provide greater geographic coverage have focused on approaches that use the data in adequately sampled areas and statistical models to extrapolate disease estimates for poorly sampled areas largely based on sociodemographic factors.[16] But many of these techniques have not been validated, and in the few instances when they have been compared, these approaches do not always work as well as expected.[6 7] Our mailed health survey found that countywide adjusted diabetes prevalence was 12.7%. This estimate is much higher than the CDC's most recent estimate of 9.5% in 2015, which is based on a modelling approach. For a given area, these modelling approaches can be especially imprecise when used to estimate disease prevalence in areas with low response rates, which includes many rural regions.

Other efforts to advance health surveillance methods have experimented with the use of claims data and electronic health records to provide estimates of disease prevalence. A recent study demonstrated that emergency claims data could be used to estimate chronic disease prevalence in New York City, and this approach was validated with results obtained from an annually performed citywide health survey. In this urban study, it was found that conditions including diabetes, hypertension, and asthma had correlations of 0.86, 0.88, and 0.77 respectively when analyzed among 34 sub-county areas.[14]

With our novel method of using emergency claims data to estimate chronic disease burden, we identified health records for a substantial majority of all adults in Sullivan County using five years of emergency claims data. Furthermore, the demographic patterns among this population of unique ED patients were much closer to Census estimates than our countywide mailed survey. Under-representation of certain demographic groups, especially minorities, is a common problem of traditional survey methods that can be adjusted for, as long as geographically matched sociodemographic data exists.[30] In rural areas where Census estimates for race and ethnicity often have wide confidence intervals, emergency claims data may provide an alternative population sample that closely mirrors the underlying population in a given region.

For some conditions such as diabetes and asthma, we found strong correlation between the two estimation methods for sub-county disease prevalence. For the other conditions studied, the strength of correlation was weaker. This may be attributable to disease-specific differences in the validity of both ED claims data and self-reported survey data. Prior research has shown that, for both data sources, validity is routinely higher for diabetes and asthma, but lower for other conditions like hyperlipidemia, with low sensitivity (i.e., under-diagnosis) being the reason for poor correlation.[8-10] It should be noted that though some conditions like COPD are a frequent primary diagnosis for a patient's ED visit, COPD may not be frequently accounted for as one of the secondary diagnoses, which are included in this ED-based surveillance approach.

Emergency claims data are already widely collected around the country, can capture a large population sample, and in some areas include address data that can be used to precisely identify where patients live. By geocoding these addresses, more precise health surveillance can provide detailed maps of disease burden. This granular level of geographic detail is important because localized hot spots of disease might otherwise be hidden as they are averaged out by neighboring areas of low disease prevalence. However, some important caveats should be understood before employing these methods. There is some variation in how accurately some hospitals capture chronic disease conditions using diagnosis codes.[31] In addition, for some parts of rural America, mail is only delivered to PO boxes, therefore the more geographically detailed maps of disease prevalence based on geocoded data may not be accurate in these regions where mail is not delivered directly. Furthermore, our study found substantial variability in prevalence estimates for conditions that may not be well captured by emergency claims data. More research may be needed to determine the best approaches for estimating disease prevalence in rural areas.

Limitations

Because surveys did not ask respondents to report household size, single-adult households are likely overrepresented. Furthermore, we did not specify a method of randomly selecting an adult in households with multiple residents, which may have contributed to bias in our sample. However, we age- and gender-standardized rates to the overall population in Sullivan County, which may partially reduce this bias. While our adjustment methods reduced age- and gender-specific non-response bias, we were unable to standardize by race and ethnicity due to the very small proportion of minorities in several areas of Sullivan County. Given that minorities often have higher rates of chronic disease and tend to have lower response rates, our mailed survey may have underestimated disease prevalence. Though the groups that frequently seek emergency care and those who respond to surveys tend to diverge, there may have been some sort of parallel bias that accounted for the correlations or disease prevalence identified in our study. Our method that used emergency claims data to estimate disease prevalence is subject to many of the limitations associated with the use of administrative data. Fidelity of coding some variables including race, ethnicity, and diagnosis codes can vary by hospital and may impact resulting disease prevalence estimates. Also, these claims data are often available about a year after they have been filed, thus there is some lag in reporting. In this study, emergency claims data were collected for 2011-2015, whereas the countywide survey was performed in 2017-2018.

CONCLUSION

We found that for select conditions, ED data may be useful for tracking disease prevalence in rural areas and may provide more geographically precise estimates. Given the infrastructure already in place to collect this data, efforts could be focused on collecting more accurate diagnosis codes and more detailed geographic data. This approach could potentially help match the limited health resources of a rural county to the geographic areas with the highest burden of disease.

Figure Legends

Figure 1. Sub-County Estimates of Adjusted Disease Prevalence Based on Mailed Survey Responses

Figure 2. Geographically Detailed Estimates of Unadjusted Disease Prevalence Based on Emergency Claims Data

Acknowledgments

The content of the study reflects the views of the authors and not the official position of the Sullivan County Public Health Department nor the NYU School of Medicine.

Author contributions

Study conception and design: DCL, LET, NAM; acquisition of the data: DCL, MO, MVN, AN, AJV; analysis and interpretation of the data: DCL, JMF, MO, CAK, CJS, AJV, LET; drafting of the manuscript: DCL; critical revision of the manuscript for intellectual content: JMF, MO, CAK, MVN, AN, CJS, AJV, LET. DCL is the guarantor of this work and had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This work was supported by the New York State Health Foundation grant number 16-04083. The funding source had no role in study design, conduct, data collection, data analysis, preparation of the manuscript, or the decision to submit the manuscript for publication.

Competing of interests

None declared.

Patient consent for publication

Not required

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination of our research.

Ethics approval

This study was approved by NYU School of Medicine's Institutional Review Board.

Data availability statement

Deidentified participant data from the countywide health survey from 2017-2018 can be made available upon reasonable request by contacting Dr. David Lee at david.lee@nyumc.org so long as the requester agrees to the following conditions: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

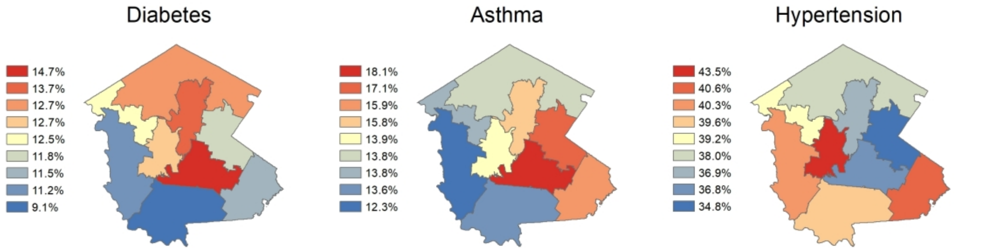
REFERENCES

1. County Health Rankings and Roadmaps. 2018 County Health Rankings Report Madison, WI: University of Wisconsin Population Health Institute; 2018 [Available from: http://www.countyhealthrankings.org/sites/default/files/state/downloads/CHR2018_NY.pdf accessed November 8th 2018.
2. Harris JK, Beatty K, Leider JP, et al. The Double Disparity Facing Rural Local Health Departments. *Annual review of public health* 2016;37:167-84. doi: 10.1146/annurev-publhealth-031914-122755 [published Online First: 2016/01/07]
3. Brundisini F, Giacomini M, DeJean D, et al. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. *Ontario health technology assessment series* 2013;13(15):1-33. [published Online First: 2013/11/15]
4. Puma JE, Belansky ES, Garcia R, et al. A Community-Engaged Approach to Collecting Rural Health Surveillance Data. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association* 2017;33(3):257-65. doi: 10.1111/jrh.12185 [published Online First: 2016/06/16]
5. Bellamy GR, Bolin JN, Gamm LD. Rural Healthy People 2010, 2020, and beyond: the need goes on. *Family & community health* 2011;34(2):182-8. doi: 10.1097/FCH.0b013e31820dea1c [published Online First: 2011/03/08]
6. Feldman JM. Re: Obesity and place: Chronic disease in the 500 largest U.S. cities. *Obesity research & clinical practice* 2018;12(5):490. doi: 10.1016/j.orcp.2018.08.003 [published Online First: 2018/09/18]
7. Wang Y, Holt JB, Zhang X, et al. Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013. *Preventing chronic disease* 2017;14:E99. doi: 10.5888/pcd14.170281 [published Online First: 2017/10/20]
8. Hebert PL, Geiss LS, Tierney EF, et al. Identifying persons with diabetes using Medicare claims data. *American journal of medical quality : the official journal of the American College of Medical Quality* 1999;14(6):270-7. doi: 10.1177/106286069901400607 [published Online First: 2000/01/07]
9. Tessier-Sherman B, Galusha D, Taiwo OA, et al. Further validation that claims data are a useful tool for epidemiologic research on hypertension. *BMC public health* 2013;13:51. doi: 10.1186/1471-2458-13-51 [published Online First: 2013/01/22]
10. Robinson JR, Young TK, Roos LL, et al. Estimating the burden of disease. Comparing administrative data and self-reports. *Medical care* 1997;35(9):932-47. [published Online First: 1997/09/23]
11. Colantonio LD, Levitan EB, Yun H, et al. Use of Medicare Claims Data for the Identification of Myocardial Infarction: The Reasons for Geographic And Racial Differences in Stroke (REGARDS) Study. *Medical care* 2018 doi: 10.1097/mlr.0000000000001004 [published Online First: 2018/10/27]
12. Hope KG, Merritt TD, Durrheim DN, et al. Evaluating the utility of emergency department syndromic surveillance for a regional public health service. *Communicable diseases intelligence quarterly report* 2010;34(3):310-8. [published Online First: 2010/11/26]

13. Centers for Disease Control and Prevention. Health, United States, 2013: With Special Feature on Prescription Drugs. Hyattsville, MD: National Center for Health Statistics, 2014.
14. Lee DC, Long JA, Wall SP, et al. Determining Chronic Disease Prevalence in Local Populations Using Emergency Department Surveillance. *American journal of public health* 2015;105(9):e67-74. doi: 10.2105/ajph.2015.302679 [published Online First: 2015/07/17]
15. Centers for Disease Control and Prevention. Public Health Surveillance in Rural Communities 2019 [Available from: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/surveillance-in-rural-communities.htm.
16. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. *American journal of preventive medicine* 2004;26(5):453-60. doi: 10.1016/j.amepre.2004.02.004 [published Online First: 2004/05/29]
17. Zhang X, Holt JB, Yun S, et al. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American journal of epidemiology* 2015;182(2):127-37. doi: 10.1093/aje/kwv002 [published Online First: 2015/05/10]
18. Coburn AF, MacKinney AC, McBride TD, et al. Choosing Rural Definitions: Implications for Health Policy Iowa City, IA: Rural Policy Research Institute Health Panel, 2007.
19. Centers for Disease Control and Prevention. SMART BRFSS MMSA Methodology Atlanta, GA2016 [Available from: https://www.cdc.gov/brfss/smart/2016/2016_SMART_BRFSS_MMSA_Methodology.pdf accessed November 8th 2018.
20. NYS ITS GIS Program Office. NYS GIS Clearinghouse Albany, NY [Available from: www.gis.ny.gov accessed November 8th 2018.
21. Mullen PD, Savas LS, Bundy LT, et al. Minimal intervention delivered by 2-1-1 information and referral specialists promotes smoke-free homes among 2-1-1 callers: a Texas generalisation trial. *Tobacco control* 2016;25(Suppl 1):i10-i18. doi: 10.1136/tobaccocontrol-2016-053045 [published Online First: 2016/10/05]
22. Pierannunzi C, Hu SS, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011. *BMC medical research methodology* 2013;13:49. doi: 10.1186/1471-2288-13-49 [published Online First: 2013/03/26]
23. Lee DC, Doran KM, Polsky D, et al. Geographic variation in the demand for emergency care: A local population-level analysis. *Healthcare (Amsterdam, Netherlands)* 2016;4(2):98-103. doi: 10.1016/j.hjdsi.2015.05.003 [published Online First: 2016/06/28]
24. Klein RJ, Schoenborn CA. Age adjustment using the 2000 projected U.S. population. Healthy People Statistical Notes. Hyattsville, Maryland: National Center for Health Statistics, January 2001.
25. American Association for Public Opinion Research. Mailed Surveys of Unnamed Persons Oakbrook Terrace, ILApril 2016 [Available from:

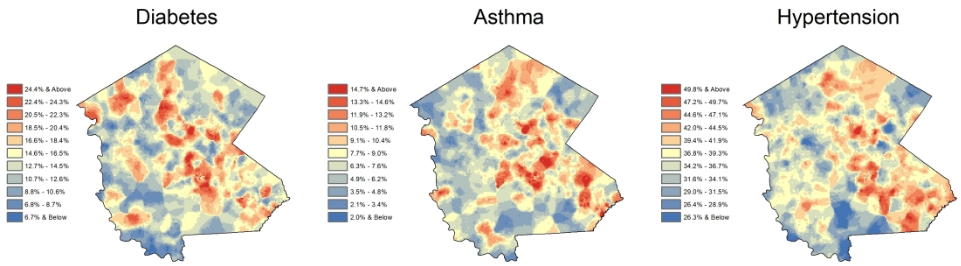
https://www.aapor.org/getattachment/Standards-Ethics/Standard-Definitions-(1)/Mail_Surveys_Unnamed_Persons_042216.pdf.aspx accessed November 8th 2018.
26. Brownson RC, Fielding JE, Green LW. Building Capacity for Evidence-Based Public Health: Reconciling the Pulls of Practice and the Push of Research. *Annual review of public health* 2018;39:27-53. doi: 10.1146/annurev-publhealth-040617-014746 [published Online First: 2017/11/23]
27. Galea S, Tracy M. Participation rates in epidemiologic studies. *Annals of epidemiology* 2007;17(9):643-53. doi: 10.1016/j.annepidem.2007.03.013 [published Online First: 2007/06/08]
28. Keeter S, Hatley N, Kennedy C, et al. What Low Response Rates Mean for Telephone Surveys Washington DC: Pew Research Center; 2017 [Available from: http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/ accessed July 25th 2017.
29. O'Connor A, Wellenius G. Rural-urban disparities in the prevalence of diabetes and coronary heart disease. *Public health* 2012;126(10):813-20. doi: 10.1016/j.puhe.2012.05.029 [published Online First: 2012/08/28]
30. Link MW, Mokdad AH, Stackhouse HF, et al. Race, ethnicity, and linguistic isolation as determinants of participation in public health surveillance surveys. *Preventing chronic disease* 2006;3(1):A09. [published Online First: 2005/12/17]
31. Lee DC, Swartz JL, Koziatek CA, et al. Assessing the Reliability of Performing Citywide Chronic Disease Surveillance Using Emergency Department Data from Sentinel Hospitals. *Population health management* 2017 doi: 10.1089/pop.2016.0168 [published Online First: 2017/03/25]

25



Sub-County Estimates of Adjusted Disease Prevalence Based on Mailed Survey Responses

228x76mm (300 x 300 DPI)



Geographically Detailed Estimates of Unadjusted Disease Prevalence Based on Emergency Claims Data

304x101mm (300 x 300 DPI)

Sullivan County Public Health Services
50 Community Lane, PO Box 590
Liberty, New York 12754
Phone: (845) 292-5910
Fax: (845) 513-2276

2017 County Wide Health Survey

The Public Health Services and the NYU School of Medicine have partnered to perform an important county wide health survey sponsored by the New York State Health Foundation to better understand health in Sullivan County. The survey takes approximately 5 to 10 minutes to complete, and you will receive a **$10** Visa gift card by mail for your time. It is completely voluntary, and your responses will be kept confidential. No information will be collected or used to identify you or any other individuals who participate in the survey. You must be a Sullivan County resident to be eligible. If there is more than one adult in your household, then choose only one person randomly. **Please return in the enclosed envelope by October 28, 2017.**

1. Is your primary home or residence located in Sullivan County? Yes No

IF YOU CIRCLED NO TO THIS QUESTION, THERE IS NO NEED TO CONTINUE.

---

2. How many months per year do you live in Sullivan County? _____ of 12 months

3. How long does it take you to get to your post office from home? _______ Minutes

4. Is this amount of time above how long it takes by: Driving Walking

5. In general, your health is: Excellent Very Good Good Fair Poor

Has a doctor, nurse, or other health professional EVER told you that you had:

6. Hypertension or High Blood Pressure: Yes No

7. High Cholesterol or Hyperlipidemia: Yes No

8. Cancer: Yes No If Yes, what type(s)? ____________________________

9. Diabetes Yes No

10. If yes, what kind of Diabetes? Type 1 Type 2 Pregnancy Related Not Sure

11. About how tall are you without shoes? _____ Feet and _____ Inches

12. About how much do you weigh without shoes? _________________ Pounds

13. Have you smoked at least 100 cigarettes in your entire life? Yes No

14. Do you currently smoke cigarettes: Every day Some days Not at all

Please turn the page over for additional questions.

15. Has a doctor, nurse or other health professional ever told you that you had asthma?

Yes No

16. During the last 12 months, have you had an episode of asthma or asthma attack?

Yes No

17. Has a doctor, nurse or other health professional ever told you that you had chronic obstructive pulmonary disease (COPD) or emphysema? Yes No

18. How many children less than 18 years of age live in your household? ____ children

19. Of these children less than 18 years old who live with you, have any of them ever been diagnosed with asthma? If yes, how many of these children? ____ children

20. Of these children less than 18 years old who live with you, have any of them ever been diagnosed with diabetes? If yes, how many of these children? ____ children

In order to provide an accurate assessment of health in Sullivan County, we need to ask the following basic demographic questions about you.

21. What is your age? 18-24 25-34 35-44 45-54 55-64 65-84 85 or older

22. Are you? Male Female

23. Are you Hispanic, Latino or Spanish in origin? Yes No

24. Which of these is your race? White Black Asian Other
(You may circle more than one.)

25. Would you like your gift card sent to the same address we used to mail this survey?

Yes No *(If you circled no, then you will need to call the phone number below to provide a different address. You will also need your survey number listed below.)*

Thank you so much for your time. Your responses will be kept confidential. If you have any questions regarding this survey, please call (845) 397-7747.

Si necesita esta encuesta en español, por favor llame al número de teléfono que aparece arriba.

Sullivan COUNTY CATSKILLS Mountains of Opportunities

Survey Number 00000-000

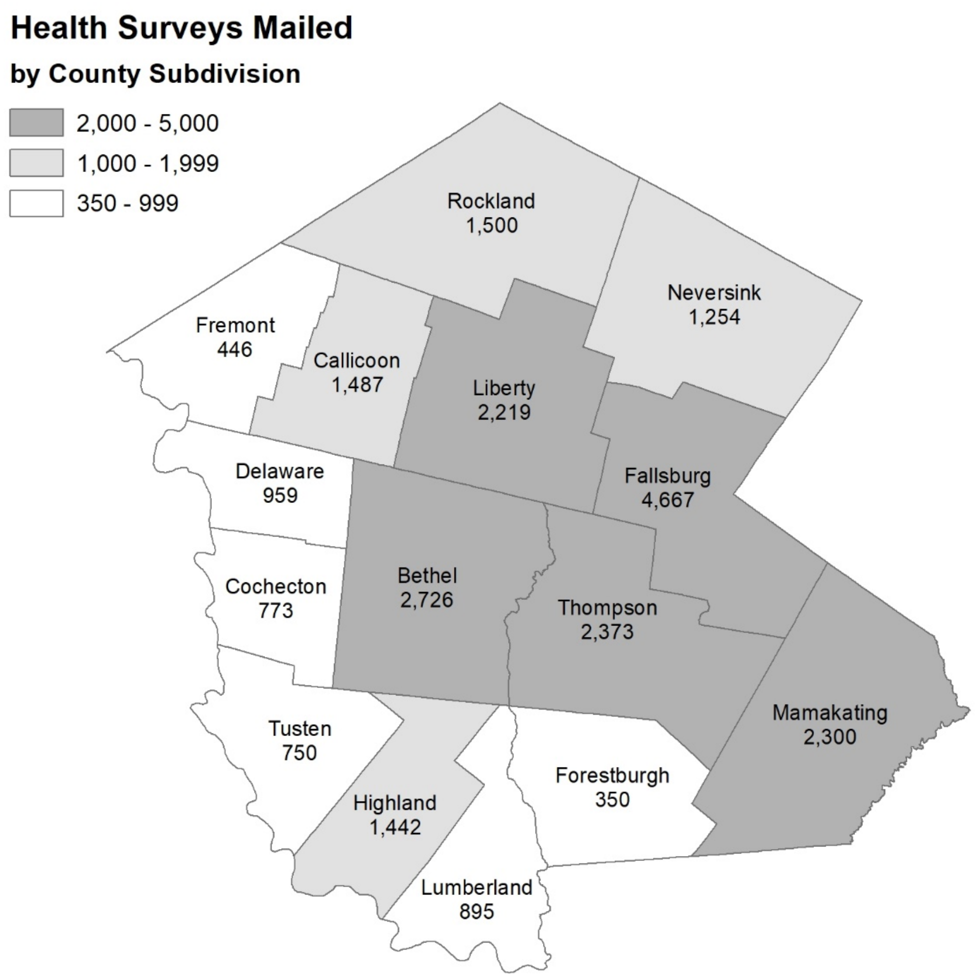
Public Health Prevent. Promote. Protect.

Supplemental Table 1: Households, Mailings, and Responses by Geographic Level in Sullivan County

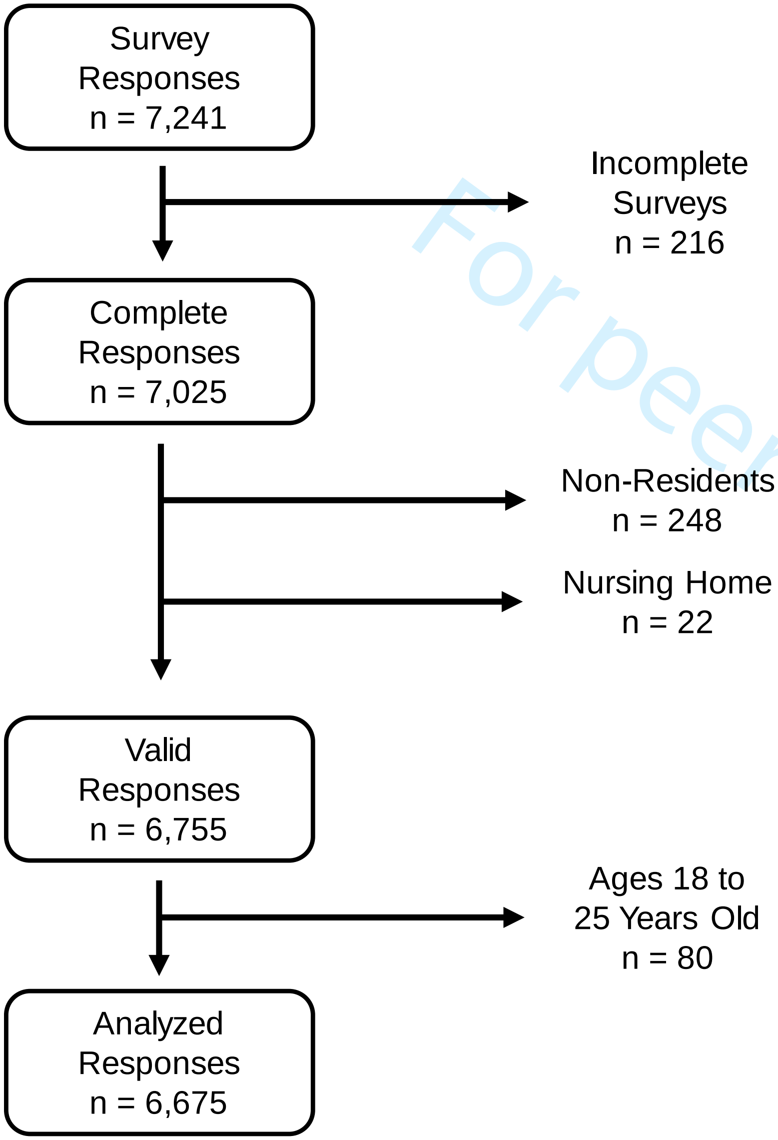
Zip Code	Population	Households	Mailings	Responses	Subdivison	Population	Households	Mailings	Responses	Subcounty	Population	Households	Mailings	Responses
12720	143	159	159	51	Bethel	2,900	2,774	2,726	719	Bethel	2,900	2,774	2,726	719
12749	266	382	382	97	Bethel					Bethel				
12762	288	323	323	104	Bethel					Bethel				
12778	437	618	618	179	Bethel					Bethel				
12783	1,396	798	750	157	Bethel					Bethel				
12786	370	494	494	131	Bethel					Bethel				
12724	217	199	199	85	Callicoon	2,175	1,731	1,487	526	Callicoon / Fremont	2,817	2,177	1,933	707
12748	1,172	994	750	205	Callicoon					Callicoon / Fremont				
12766	309	179	179	68	Callicoon					Callicoon / Fremont				
12791	477	359	359	168	Callicoon					Callicoon / Fremont				
12736	89	84	84	27	Fremont	642	446	446	181	Callicoon / Fremont				
12741	157	159	159	69	Fremont					Callicoon / Fremont				
12760	353	152	152	56	Fremont					Callicoon / Fremont				
12767	43	51	51	29	Fremont					Callicoon / Fremont				
12726	959	551	551	186	Cochecton	1,142	773	773	248	Cochecton / Delaware / Tusten	4,139	2,842	2,482	763
12752	183	222	222	62	Cochecton					Cochecton / Delaware / Tusten				
12723	1,391	893	750	212	Delaware	1,711	1,102	959	289	Cochecton / Delaware / Tusten				
12745	127	92	92	36	Delaware					Cochecton / Delaware / Tusten				
12750	193	117	117	41	Delaware					Cochecton / Delaware / Tusten				
12764	1,286	967	750	226	Tusten	1,286	967	750	226	Cochecton / Delaware / Tusten				
12733	1,070	550	550	79	Fallsburg	8,849	5,737	4,667	966	Fallsburg	8,849	5,737	4,667	966
12738	327	132	132	34	Fallsburg					Fallsburg				
12747	1,143	735	735	170	Fallsburg					Fallsburg				
12759	886	574	574	132	Fallsburg					Fallsburg				
12763	409	426	426	129	Fallsburg					Fallsburg				
12779	1,244	1,162	750	114	Fallsburg					Fallsburg				
12788	2,381	1,005	750	180	Fallsburg					Fallsburg				
12789	1,389	1,153	750	128	Fallsburg					Fallsburg				
12729	82	41	41	16	Forestburgh	781	350	350	122	Forestburgh /. Highland / Lumberland	4,314	2,763	2,687	856
12777	699	309	309	106	Forestburgh					Forestburgh /. Highland / Lumberland				
12719	1,042	520	520	179	Highland	2,013	1,442	1,442	493	Forestburgh /. Highland / Lumberland				
12732	577	480	480	155	Highland					Forestburgh /. Highland / Lumberland				
12743	242	184	184	66	Highland					Forestburgh /. Highland / Lumberland				
12792	152	258	258	93	Highland					Forestburgh /. Highland / Lumberland				
12737	1,382	826	750	193	Lumberland	1,520	971	895	241	Forestburgh /. Highland / Lumberland				
12770	138	145	145	48	Lumberland					Forestburgh /. Highland / Lumberland				
12734	697	652	652	162	Liberty	6,119	4,875	2,219	620	Liberty	6,119	4,875	2,219	620
12754	4,702	3,406	750	178	Liberty					Liberty				
12768	532	556	556	175	Liberty					Liberty				
12787	188	261	261	105	Liberty					Liberty				
12721	4,153	2,108	750	166	Mamakating	8,519	5,368	2,300	578	Mamakating	8,519	5,368	2,300	578
12722	77	200	200	59	Mamakating					Mamakating				
12769	156	100	100	40	Mamakating					Mamakating				
12781	211	133	133	42	Mamakating					Mamakating				
12785	704	367	367	63	Mamakating					Mamakating				
12790	3,218	2,460	750	208	Mamakating					Mamakating				
12725	143	103	103	24	Neversink	2,046	1,286	1,254	400	Neversink / Rockland	6,736	4,134	2,754	789
12740	1,208	782	750	236	Neversink					Neversink / Rockland				
12765	695	401	401	140	Neversink					Neversink / Rockland				
12758	2,966	1,997	750	204	Rockland	4,690	2,848	1,500	389	Neversink / Rockland				
12776	1,724	851	750	185	Rockland					Neversink / Rockland				
12701	7,795	6,044	750	228	Thompson	10,104	8,414	2,373	677	Thompson	10,104	8,414	2,373	677
12742	62	197	197	66	Thompson					Thompson				
12751	561	564	564	174	Thompson					Thompson				
12775	1,648	1,497	750	170	Thompson					Thompson				
12784	38	112	112	39	Thompson					Thompson				
Total	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**	**Total**	**54,497**	**39,084**	**24,141**	**6,675**

Note: Population estimates are based on ZCTAs and may not align with ZIP codes. A population adjustment was required in ZIP code 12729 due to gross mismatch between ACS estimate and number of households from sampling frame.

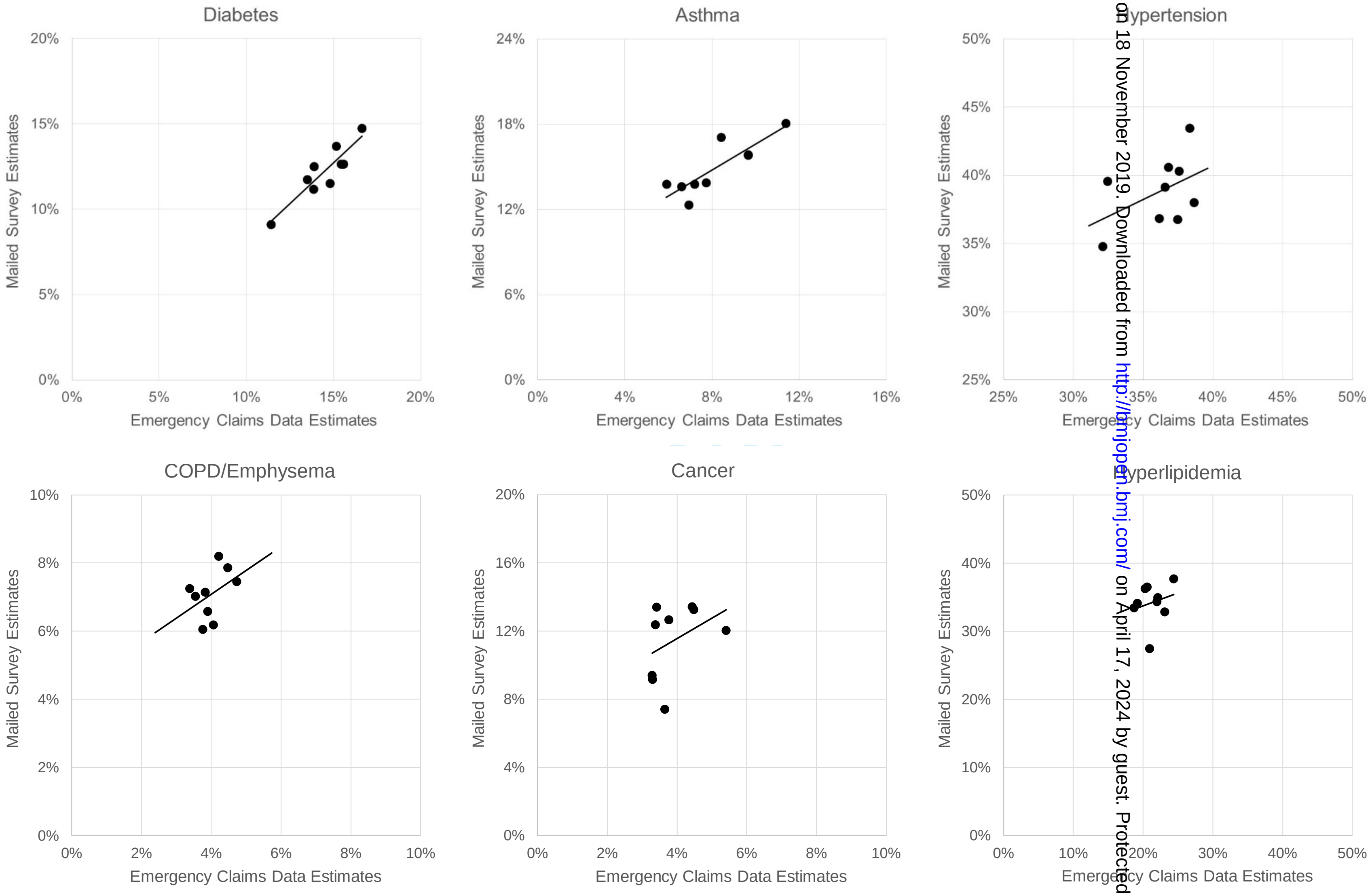
Supplemental Figure 1: Health Survey Mailed by County Subdivision



Supplemental Figure 2: Diagram of Survey Responses and Exclusions



Supplemental Figure 3: Comparison of Sub-County Estimates of Disease Prevalence



STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **Page 1**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **Pages 2-3**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **Pages 4-5**
Objectives	3	State specific objectives, including any prespecified hypotheses **Page 5, Paragraph 2**
Methods		
Study design	4	Present key elements of study design early in the paper **Page 6, Paragraph 1**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **Pages 6-7**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants **Pages 6-7**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **Page 8, Paragraph 1**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **Pages 8-9**
Bias	9	Describe any efforts to address potential sources of bias **Pages 8-9**
Study size	10	Explain how the study size was arrived at **Page 11**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **Pages 8-9**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **Pages 8-10**
		(*b*) Describe any methods used to examine subgroups and interactions **N/A**
		(*c*) Explain how missing data were addressed **Page 11**
		(*d*) If applicable, describe analytical methods taking account of sampling strategy **Pages 8-9**
		(*e*) Describe any sensitivity analyses **N/A**
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **Page 11**
		(b) Give reasons for non-participation at each stage **Page 11**
		(c) Consider use of a flow diagram **Supplemental Materials**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **Page 11, Paragraph 2**
		(b) Indicate number of participants with missing data for each variable of interest **Page 11, Supplemental Materials**
Outcome data	15*	Report numbers of outcome events or summary measures **Page 13**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and

		their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **Page 13**
		(*b*) Report category boundaries when continuous variables were categorized **N/A**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period **N/A**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **Page 14**
Discussion		
Key results	18	Summarise key results with reference to study objectives **Page 15, Paragraph 1**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **Page 18, Paragraph 1**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **Pages 15-17**
Generalisability	21	Discuss the generalisability (external validity) of the study results **Page 18**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **Page 20**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.