Executive Overview

Generating Sub-County Health Data Products: Methods and Recommendations from a Multi-State Pilot Initiative

Background:

In 2010, the County Health Rankings & Roadmaps (CHR&R) program achieved an important milestone of providing overall health measures for nearly every county nationwide. However, after several years of producing the rankings, limitations of county-level data emerged as challenges for local public health planning (e.g., county-level rankings mask gaps among sub-county population and geographies). A call out was subsequently issued to public health researchers to enhance sub-county data availability.

In 2015, CHR&R funded pilot work by state departments of health, universities, and hospital association partnerships in New York, California, and Missouri to explore ways to build data infrastructures that enhance local data availability, and develop sub-county health measures compatible with CHR&R. The overall aims of the pilot projects were to: 1) provide data to support local community health needs assessments and development of community health improvement plans; and 2) develop analytical capability for small area data analyses and presentation to support public health activities.

Since the conclusion of the pilot projects, the research grantees have been working to develop both a **published, peer-review manuscript** and a **supplemental companion white paper** summarizing key thematic considerations, lessons learned and helpful takeaways from the three pilot projects to support public health practitioners in responding to the growing need and demand for sub-county health data. Areas off emphasis for the manuscript and white paper include:

1. **Conceptual development for data sources and measures**
2. **Analyzing and presenting small-area and sub-population measures for public health, healthcare, and lay audiences**
3. **Positioning sub-county data initiatives for growth and sustainability**

Intended audiences include state and local public health program managers, data analysts, surveillance staff, and others who would be interested in conducting similar projects of their own.

The manuscript and white paper will jointly summarize key considerations and lessons learned that were carefully selected from the three pilot projects. The considerations and lessons learned will be organized thematically with narrative supported by a tabular presentation:

Information and Guidance:

Conceptual Development

Building a valid, reliable and sustainable measure set at the sub-county level begins with establishing a solid conceptual foundation. Key considerations covered in this section are presented to help readers think through critical foundational elements including:

a) Clearly defining the right target audience
b) Systematically selecting a set of measures that will sustainably fulfill identified needs
c) Weighing technical and practical considerations involved with defining the most useful sub-county geographic unit and time aggregation for reporting of results

Data Analysis and Presentation

Developing and implementing a methodologically sound process for producing and reporting small area health indicator estimates can be complicated. The largest section of the manuscript will provide an overview of common experiences and challenges with technical aspects of deriving and reporting sub-county small area health measures with specific attention to:

d) Applying analytic methods to generate empirically sound sub-county estimates
e) Applying suppression criteria according to requirements of specific data sources to protect individual confidentiality
f) Assessing data stability and planning to indicate or flag the unstable estimates in publications or data products
g) Designing effective tabular and visual presentations of results for targeted users
h) Putting processes in place to automate production
i) Implementing an effective strategy and mechanism for disseminating results

Positioning for Growth and Sustainability

Supporting community action typically requires translation of initial work to ongoing delivery of sub-county data and reporting results over time. Key considerations considered in this section will include

j) Engaging targeted stakeholders early in the development process
k) Planning and budgeting for sustained operations
l) Securing ongoing funding support

The manuscript content will conclude with a summary of common challenges, lessons learned, and takeaway recommendations for potential sub-county measure developers.

The supplemental white paper serve as a companion reference to the manuscript, and will be published online as a home page that links to the published manuscript, includes a brief summary of each project, embedded links to more detailed project descriptions, and supplemental reference materials from each project. Supplemental materials will include useful documents such as data dictionaries, user training presentations, public use data files, links to public-facing reports, and statistical programming code used to produce various sub-county estimates.

Appendix:

Supplemental White Paper - Analyzing and Presenting Small-area and Sub-population Data

Introduction

This report is a collection three white papers describing methods and results from three research projects in a multistate (California, Missouri, New York) sub-county health data pilot initiative. The white papers are intended to be helpful resources for researchers, practitioners, and communities who are interested in carrying out similar work.

Aims and objectives

1. To share experience from three research projects that used multiple data sources to analyze and generate sub-county level (small area) data for health-related measures aligned with the County Health Rankings and Roadmaps model, to support community health needs assessment, disparity identification, and targeted intervention.
2. To discuss sub-county level data analysis for different types of data sources, including individual-level count data (e.g., birth, hospitalization, mortality) and individual-level survey data (e.g., American Community Survey, expanded Behavioral Risk Factor Surveillance System Survey).
3. To describe applications of various statistical methods for analyzing and presenting data at below-county levels.
4. To provide examples of statistical issues (e.g., confidentiality, reliability and stability) that are inherent to small-area data analysis, as well as trade-offs and rationales to consider when the issues are encountered.
5. To describe how various forms of collaboration, and solicited feedback from key stakeholders, can improve end-users' understanding and utilization of sub-county level data.

Intended audiences

- Data analysts
- Surveillance staff
- Managers of analyst staff
- Funders

Pilot Projects and Technical Information

- **California**

  In 2015, the Healthy Communities Data and Indicators Project (HCI), Office of Health Equity (OHE), California Department of Public Health (CDPH), collaborated with the County Health Rankings and Roadmaps (CHR&R) for a pilot project to develop sub-county health measures aligned with the Rankings framework. The aims of the pilot project were to generate a subset of the measures in the CHR&R model at the sub-county level (city, census tract) disaggregated by demographic groups (sex, race/ethnicity, disability status, poverty status), for California and other U.S. states. Another aim was to develop programming code to automate the downloading of source data and the generation of datasets. This white paper describes methodological details of the pilot project, especially data source selection and tradeoffs in the use of various data products. The HCI project already had three years of experience building a standardized set of statistical measures, data files, and tools for planning healthy and equitable communities in California. Data generated by the HCI helps in the assessment of the health and equity status of communities in California; the Office of Health Equity has a mandate to report on the social determinants of health to the people and Legislature of California. A white paper is available starting on page 5. Datasets are made publicly available and the complete list of measures currently available can be found here: https://www.cdph.ca.gov/Programs/OHE/Pages/HCI-Search.aspx#.

- **Missouri**
  The aim of the Missouri ZIP Health Rankings Project was to enable statewide community health improvement stakeholders to better target scarce health improvement resources to sub-county areas with greatest need by extending the established County Health Rankings measurement model to the ZIP Code-level. Measure derivation for the project involved developing a broad set of ZIP code-level candidate measures statewide hospital encounter and census databases, using advanced analytic modeling methods to derive analog composite score for County Health Rankings measurement domains and generating ZIP Code-level rankings based on derived measures for 976 Missouri ZIP codes. Following completion of the project, Missouri ZIP Health Rankings data and measures were published on www.exploreMOhealth.org, an online interactive portal developed to in partnership between Missouri Hospital Association and the Missouri Foundation for Health. A full description of project results published in the Journal of Public Health Management and Practice is available here. Additional supporting documents describing the project, associated data and supporting documents used to guide report development can be found here.

- **New York**
  The New York Team analyzed multiple data sources (including births, deaths, hospitalizations, and Behavioral Risk Factor Surveillance System survey) to generate results below county level for eleven measures. These sub-county measures aligned with selected county-level measures on the County Health Rankings. The team developed 62 individual county reports in PDF format for all counties in New York State, which included county maps, tables, and graphs. These products were widely used for community health needs assessment. Full descriptions of the project, along with related materials and supporting documents for developing these reports can be found on page 36.

California CHR&R Pilot Project

Introduction

In 2015, the Healthy Communities Data and Indicators Project (HCI), Office of Health Equity (OHE), California Department of Public Health (CDPH), partnered with the County Health Rankings and Roadmaps (CHR&R) for a pilot project to develop sub-county health measures aligned with the Rankings framework. The aims of the pilot project were to generate a subset of the measures in the CHR&R model at the sub-county level (city, census tract) disaggregated by demographic groups (sex, race/ethnicity, disability status, poverty status), for California and other U.S. states. Another aim was to develop programming code to automate the downloading of source data and the generation of datasets. This white paper describes methodological details of the pilot project, especially data source selection and tradeoffs in the use of various data products. The HCI project already had three years of experience building a standardized set of statistical measures, data files, and tools for planning healthy and equitable communities in California. Data generated by the HCI helps in the assessment of the health and equity status of communities in California; the Office of Health Equity has a mandate to report on the social determinants of health to the people and Legislature of California. Datasets are made publicly available and the complete list of measures currently available can be found here: https://www.cdph.ca.gov/Programs/OHE/Pages/HCI-Search.aspx#.

Project Aims

1. To generate datasets for a subset of the measures in the County Health Rankings Roadmaps (CHR&R) model, geographically disaggregated to the sub-county level.
2. To identify data sources that further disaggregate the measures by demographic characteristics (sex, race/ethnicity, disability status, and poverty status).
3. Develop programming code to automate the downloading of source data and the generation of sub-county measure datasets.

Intended Audience

The people of California, the California Legislature, local public health agencies, California state agencies participating in the Health in All Policies Task Force, the State health improvement plan “Let’s Get Healthy California”, other health organizations, community-based organizations, and researchers.

Intended Uses of the Data

The goal of the Office of Health Equity (OHE), California Department of Public Health (CDPH), through its Healthy Communities Data and Indicators Project (HCI), is to provide a standardized

set of statistical measures, data files, and tools for planning healthy and equitable communities in California. The data helps California assess the health and equity status of communities and informs the mandatory reporting on the social determinants of health from the Office of Health Equity to the California Legislature. The HCI datasets are publicly available. Local health departments and researchers could use the datasets in their reporting or adapt the methods from this project to create indicator datasets for their jurisdictions or projects.

Methods

The HCI project has been producing datasets for indicators or measures of the social determinants of health since 2012. The framework for the HCI is the definition of "What is a Healthy Community?" from the California Health in All Policies Task Force (Figure 1). Some of the methods and standards developed for the HCI were used for this pilot project. The research and development phase of the HCI occurred during the years 2012–2014. During this time, stakeholder engagement including with local health departments and other state agencies, was conducted and feedback was obtained regarding indicator definitions, data sources, data presentation standards, and metadata standards. The complete list of indicators currently used by the HCI is found here: https://www.cdph.ca.gov/Programs/OHE/Pages/HCI.aspx.

Figure 1. "What is a Healthy Community?" definition from the California Health in All Policies Task Force. (Source: California Health in All Policies Task Force. Health in All Policies Task Force Report to the Strategic Growth Council. Retrieved from http://www.sgc.ca.gov/programs/hiap/docs/2010-HiAP_Task_Force_Report-_Dec_2010.pdf)

The HCI uses the following criteria to select indicators:

- Validity
    - Measures what it purports to measure
    - Evidence linking indicator to health outcomes

- Technical and Data Properties
    - Data source(s) owned and collected by a recognized organization
    - Timeliness (time lag and frequency of updates)
    - Data quality (completeness, missing data, accuracy)
    - Variety of geographic levels available, including Census tract
    - Administrative accessibility (public domain, proprietary, confidentiality, costs)
    - Current use and acceptability to stakeholders
    - Straightforward mechanics of data collection, aggregation, and reporting

- Usable and Understandable to Users

Environmental Scan of Data Sources

For this pilot project, ten measures from the CHR&R model were identified and prioritized in partnership with CHR&R staff:

- **Child Poverty:** Percentage of children under age 18 in poverty
- **Income Inequality:** Ratio of household income at the 80th percentile to income at the 20th percentile
- **Driving Alone to Work:** Percentage of the workforce that drives alone to work
- **Some Post-Secondary Education:** Percentage of adults ages 25-44 years with some post-secondary education
- **Unemployment:** Percentage of population ages 16 and older unemployed but seeking work
- **Housing Problems:** Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities
- **Health Status:** Percentage of adults reporting fair or poor health
- **Adult body mass index:** Percentage of adults that report a BMI of 30 or more
- **Insurance Status:** Percentage of population under age 65 without health insurance
- **Violent Crime:** Number of reported violent crime offenses per 100,000 population

Four of the selected CHR&R measures overlapped with existing HCI measures: child poverty, unemployment, health insurance, and violent crime; the data sources for these measures had

been identified prior to the start of the pilot project. Data sources were researched for the remaining six measures. For three of the remaining measures (income inequality, driving alone to work, and some post-secondary education) data was available from the American Community Survey (ACS) from the U.S. Census Bureau. Data for the housing problems measure was available from the U.S. Department of Housing and Urban Development (HUD).

The University of California Los Angeles, California Health Interview Survey Neighborhood Edition (AskCHIS NE, http://askchisne.ucla.edu/) was identified as a data source to provide small area modeled estimates (SAE) for sub-county level geographies for three health measures (health status, adult body mass index, and insurance status). The SAE models use primary data from the California Health Interview Survey (CHIS, http://healthpolicy.ucla.edu/CHIS/Pages/default.aspx), the largest health survey in the country which provides important information on the health, health behaviors and access to health care services of Californians.

Obtaining Data

Following the identification of sources, data was obtained by either direct request to the source as was the case for the Federal Bureau of Investigation (FBI); or by using an Application Programming Interface (API) to download the data, in the case of the American Community Survey (ACS). Three of the datasets were obtained via a sub-award contract and a data user agreement with the University of California, Los Angeles, California Health Interview Survey Neighborhood Edition (AskCHIS NE). Data from HUD was downloaded directly from their Comprehensive Housing Affordability Strategy (CHAS) website.

SAS or R code was developed for data imports (or downloads in the case of API), manipulation, aggregation and statistical calculations for the FBI, HUD and ACS data. One of the goals of the pilot projects was to automate data downloading of sub-county measures datasets, in contrast to manual downloading. CHR&R was interested in finding ways to automate the calculations for a large number of counties and sub county geographies (potentially for all of the United States). Thus, priority was given to take advantage of ACS table products that were available and easily downloadable via the Census API at the time of the pilot project (Fall 2015-Spring 2016). It was decided to prioritize use of datasets available via API as a source; in some cases, this resulted in adopting datasets that somewhat departed from the original CHR&R definition (usually in relation to the population universe) and using ACS tables that required aggregating multiple sub-categories to generate approximate standard errors. A more detailed explanation of these limitations can be found in the section Measure Development Using ACS Data Example. The “acs” package for R software (https://cran.r-project.org/web/packages/acs/acs.pdf) was used to obtain ACS data via U.S. Census API.

Geographic Units and Time Aggregation

This project aimed to use the smallest geographical units available at the sub-county level. This was dependent on the data source and there was no uniformity on the geographic units available, which included city, ZIP Code Tabulation Area (ZCTA), and census tract (Table 1).

In order to provide geographically disaggregated and/or demographically disaggregated data, it is necessary to either suppress data derived from small counts (see Data Suppression below) or aggregate data for multiple years. Survey data at the census tract level from both the ACS and the HUD-CHAS data is already provided in 5-year aggregates. The modeled data at the ZCTA level from AskCHIS NE uses data from a two-year survey cycle. The annual count data from the FBI used to generate the violent crime rate was not aggregated over time.

Data Suppression

All sources apply their own internal data suppression criteria before release. The ACS requires at least 7,000 people in the specific population subgroup for its 5-year aggregate releases. More information on ACS data suppression can be found here: https://www2.census.gov/programs-surveys/acs/tech_docs/data_suppression/ACSO_Data_Suppression.pdf. The HUD-CHAS dataset is derived from the ACS and applies the same suppression rules (https://www.huduser.gov/portal/datasets/cp/CHAS/data_doc_chas.html). The FBI data quality guidelines indicate that all UCR data is reviewed for reasonableness before it disseminated, however there is no clear reference to any suppression guidelines (https://ucr.fbi.gov/data-quality-guidelines-new).

AskCHIS NE implemented data suppression using their program guidelines before releasing the data to the pilot project. The Stability and Pooling methodology and suppression criteria for the AskCHIS NE project are reproduced below:

"The coefficient of variation (CV) was calculated for each estimate to assess statistical stability. The coefficient of variation is defined as the ratio between the standard error of the point estimate and the point estimate. A point estimate with CV ≥ 30% is considered unstable. Unstable estimates and estimates for areas with a population universe of less than 1,000 are suppressed.

For unstable estimates, or estimates for areas with a population universe of less than 1,000, geographic locations may be combined to produce stable estimates or to achieve a sufficiently large population. The pooled point estimate and variance are population-weighted averages of the original point and variance estimates. The confidence intervals and coefficient of variations are adjusted accordingly." (Source: AskCHIS NE methods, opening free account is needed)

Given that data used in this pilot was already publicly available or it was modeled data, no privacy concerns were identified. The standard error, confidence intervals, and relative standard error (RSE) were calculated for all measures to assess reliability. Although we considered an RSE ≥ 30% as an indication of an unreliable measure, data suppression was not implemented for the measures generated from the ACS, HUD and FBI data. Based on input from stakeholder groups, when data is available it should not be suppressed due to unreliability but only flagged to avoid excluding smaller communities or populations.

Table 1. Measures Selected for the Pilot Project.

Measure	Smallest Geographic Resolution Used	Year	Demographic Strata	Source and Universe
Child Poverty: Percentage of children under age 18 in poverty	Census tract	2010-2014	Race/ethnicity	American Community Survey, U.S. Census (ACS) Table B17020: Population for whom poverty status is determined
			Sex	ACS Table B17001: Population for whom poverty status is determined
Income Inequality: Ratio of household income at the 80th percentile to income at the 20th percentile	Census tract	2010-2014	None	ACS Table B19080: Households
Driving Alone to Work: Percentage of the workforce that drives alone to work	Census tract	2010-2014	Race/ethnicity	ACS Table B08105: Workers 16 years and over
			Sex	ACS Table B08006: Workers 16 years and over
			Poverty level	ACS Table B08122: Workers 16 years and over for whom poverty status is determined
Some Post-Secondary Education: Percentage of adults ages 25-44 years with some post-secondary education	Census tract	2010-2014 for sex strata; 2014 for race/ethnicity strata	Race/ethnicity	ACS Table B15002: Population 25 years and over
			Sex	*ACS Table B15001: Population 25-44 years*
Unemployment: Percentage of population ages 16 and older unemployed but seeking work	Census tract	2010-2014	Race/ethnicity	ACS Tables C23002, B through I: Population 16 years and over
			Sex and poverty	ACS Table B17005: Civilian population 16 years and over for whom poverty status is determined.
			Disability	*ACS Table C18120: Civilian noninstitutionalized population 18 to 64 years.*
Housing Problems: Percentage of households with at least 1 of 4	Census tract	2009-2013	Race/ethnicity	U.S. Department of Housing and Urban Development

housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities				(HUD), Comprehensive Housing Affordability Strategy (CHAS), Table 2: Households (for which all of the problems were determined)
Health Status: Percentage of adults reporting fair or poor health	ZCTA (modeled)	2014 (2013-2014)	None	California Health Interview Survey Neighborhood Edition (AskCHIS NE): Adults 18-64 and 65+
Adult BMI: Percentage of adults that report a BMI of 30 or more	ZCTA (modeled)	2014 (2013-2014)	Sex, race/ethnicity	AskCHIS NE: Adults 18+
Insurance Status: Percentage of population under age 65 without health insurance	ZCTA (modeled)	2014 (2013-2014)	None	AskCHIS NE: Adults 18-64, children and teens 0-17
Violent Crime: Number of reported violent crime offenses per 100,000 population	City/town	2013	None	Federal Bureau of Investigation, Uniformed Crime Reports by County file

Measure Development Using ACS Data Example 1: Percentage of Population Ages 16 and Older Unemployed but Seeking Work

The unemployment indicator will be used to illustrate steps and decision making for the use of ACS data to obtain estimates at the county and sub-county levels. This indicator was selected as an example because it can be broken down into four demographic strata: sex, disability status, poverty level, and race and ethnicity, and there are multiple ACS products available with unemployment data.

The Local Area Unemployment Statistics (LAUS) program of the Bureau of Labor Statistics is the official source of unemployment data for the United States; it publishes modeled data for cities and towns (population 25,000 or above), but LAUS does not provide any demographic stratification. The U.S. Census Bureau's American Community Survey (ACS) is not the official source of unemployment data, but it does collect unemployment information that is available for Census tracts and cities and towns, with demographic stratification. Information about differences in the methods to estimate unemployment between the Bureau of Labor Statistics-Local Area Unemployment Statistics and the Census' American Community Survey can be found here: http://www.bls.gov/lau/acsqa.htm.

There are multiple ACS tables with different population universe definitions that cover the topics of employment by poverty, sex, disability status and race and ethnicity. The universe for the CHR&R measure is the population ages 16 and older. Table 2 compares some of the tables available to obtain demographic stratification by disability status, sex, and poverty status, and their advantages and disadvantages in the context of the pilot project.

When using ACS products to report on point estimates, tables that provide a percent estimate and its standard error are preferred over tables that require an approximation of the percent estimate by aggregating counts to produce the numerator and denominator, calculating a percent and its standard error. However, some of the preferred tables for unemployment were not available via Census API for automatic downloads at the time of the project (Fall of 2015). As shown in Table 2, ACS table S2301 provided pre-calculated percent unemployment estimate and standard error, while the other tables provided counts from where the numerator and denominator could be obtained and the percent unemployment estimate and its standard error could be approximated. When using a table for which approximation of the percent estimate was necessary, the ACS table that required the aggregation over the *least* number of categories was preferred. This was because approximation usually leads to either over- or under-estimation of the standard error of the percent and a large number of categories can exacerbate this issue (learn more at Accuracy of the Data). The example in Figure 2 illustrates this principle.

Table A and Table B both contain counts of the number of individuals 16 years and over not in the labor force by age group. To calculate the number of adults 16 years or older not in the labor force (numerator) using Table A, an aggregation over three age groups would need to be conducted. For Table B, the aggregation would need to be over five age groups. Table A is thus preferable.

Table A	Table B
Number of individuals 16 years and over	Number of individuals 16 years and over
Not in the labor force	Not in the labor force
16-20 years of age	16-20 years of age
21-30	21-30
31 or older	31-40
	41-50
	51 or older

Figure 2. Example comparing two scenarios of count data aggregation over multiple categories.

Generally, the criteria used for selection of an ACS table for the CHR&R pilot project were: (a) availability of the table via Census API to automate dataset production; (b) requires the least number of categories to be aggregated to calculate an approximate standard error; (c) table for which the universe closely matched the CHR&R definition; (d) 5-year aggregation table product, as these are the only ones with census tract level data.

ACS table S2301 was ideal to obtain the percent unemployment estimates for most of the strata. However S2301 was not available via API during the pilot project period. Consequently, table C18120 was selected to report on the disability strata because, even though the universe didn't exactly match the desired age range (18-64 years), the table was available via API and only two categories had to be aggregated to calculate the standard error of the denominator using the approximate method. Annex I presents the code that was used to download the disability data and to calculate the percent unemployment estimate and its standard error, numerator and denominator.

Table B17005 was selected to report unemployment by sex and poverty status given that the age range closely matched the range desired for the measure, and only four strata had to be aggregated to calculate the approximate standard error, as opposed to thirteen for table B23001, which overestimated the approximate standard error (Table 2; SE=0.05 versus SE=0.20). See the ACS Example 2 section below for a more detailed comparison between pre-calculated and approximated standard errors. The examples presented in Table 2 use state-level data but the same ACS tables are available for sub-county geographies. Annex II presents examples of the calculations needed to approximate all of the estimates presented in Table 2.

The ACS also provides multiple tables from which unemployment data by race and ethnicity can be obtained. The ACS published a series of tables titled B23002 "Sex by age by employment status for the population 16 years and over" that contain unemployment count estimates for "races alone or in combination" as follows:

- B23002A: White Alone
- B23002B: Black or African American Alone
- B23002C: Native American and Alaska Native Alone
- B23002D: Asian Alone
- B23002E: Native Hawaiian and Other Pacific Islander Alone
- B23002F: Some Other Race Alone
- B23002G: Two or More Races
- B23002H: White Alone, not Hispanic or Latino
- B23002I: Hispanic or Latino

The detailed B23002 tables present employed and unemployed counts for males and females, broken down by six age categories. A related collapsed table, table C23002 (A through I), presents the same data of sex by age, however, only two age categories are provided: 16-64 and 65 and older. The collapsed C23002 tables are thus preferable, given that only two age categories need to be aggregated to match the data to the universe required by the CHR&R: ages 16 and older. Both B and C tables are available via API, so automation is possible. However, a limitation is that, with the exception of the White population, individuals that

identify as Hispanic or Latino are included in the other six race groups. For public health planning purposes it is usually preferable to have data for all races separate from the Hispanic or Latino ethnicity. For this pilot project, tables C23002 were selected as the data source to report on unemployment by race and ethnicity, specifically tables C23002B through C23002I, to at least report White Alone, not Hispanic or Latino.

The ACS does produce 5-year Selected Population Tables in which the race groups exclude the Hispanic or Latino population. However, the Selected Population Tables are only available every five years, thus there are only two datasets to date (ACS started in 2005) for years 2006-2010 and 2011-2015 (which became available after the end of the pilot project). If ACS continues producing these Selected Population Tables, the next 5-year dataset will be available by 2021, therefore one limitation of this product is the time gap between data releases.

In summary, the American Community Survey offers multiple table products that can provide similar information for the construction of a measure. Multiple criteria would need to be considered to determine the best data source for a particular project depending on its goals.

Table 2. Comparison of ACS tables that cover the topics of unemployment by disability, sex and poverty: universe, unemployment percent estimate and standard error (SE), California, 2010-2014.

ACS Table	Population in the Universe	CA Unemployment Estimate	SE	Advantage for pilot project	Disadvantage for pilot project
Disability					
S2301	16 years and over with any disability	20.0%	0.24	- Matches CHR&R universe **- Percent unemployment estimate and standard error (calculated from margin of error) provided by Census**	- Table not available from Census API, manual download of geography tables (tract, city, county, state) is necessary using Census Fact Finder or other (like Census FTP)
C18120 (selected for pilot project)	Civilian noninstitutionalized 18-64 years with any disability	20.5%	0.23	- Available from Census API, which facilitates automation of data extraction for multiple geographies	- Does not match CHR&R universe - Percent unemployment estimate and standard error needs to be approximated→ denominator needs to be approximated aggregating 2 categories, numerator is available
B23024	Civilian 20-64 years for whom poverty status is determined with any disability	20.0%	0.23	- Available from Census API	- Does not match CHR&R universe - Percent unemployment estimate and standard error needs to be approximated→ denominator needs to be approximated aggregating 4 categories, numerator needs to be approximated aggregating 2 categories
Sex (Male)					
B23001	16 years and over	12.0%	0.20	- Available from Census API - Matches CHR&R universe	-Percent unemployment estimate and their standard errors need to be approximated→ denominator is available numerator needs to be approximated aggregating 13 categories
S2301	20-64 years	10.9%	0.06	**- Percent unemployment estimate and standard error (calculated from margin of error) provided by Census**	- Does not match CHR&R universe - Table not available from Census API, manual download of geography tables (tract, city, county, state) is necessary using Census Fact Finder or other (like Census FTP)
B17005 (selected)	16 years and over for whom poverty status is determined	11.7%	0.05	- Available from Census API - Somewhat matches CHR&R universe	- Percent unemployment estimate and standard error needs to be approximated→ denominator needs to be approximated aggregating 2 categories, numerator needs to be approximated aggregating 2 categories
Poverty					
S2301	16 years and over with poverty status below the poverty level in the past 12 months	31.7%	0.18	- Somewhat matches CHR&R universe **- Percent unemployment estimate and standard error (calculated from margin of error) provided by Census**	- Table not available from Census API, manual download of geography tables (tract, city, county, state) is necessary using Census Fact Finder or other (like Census FTP)
B17005 (selected)	16 years and over for whom poverty status is determined, income in the past 12 months below the poverty level	32.8%	0.18	- Available from Census API - Somewhat matches CHR&R universe	- Percent unemployment estimate and standard error needs to be approximated→ denominator needs to be approximated aggregating 2 categories, numerator needs to be approximated aggregating 2 categories

Measure Development Using ACS Data Example 2: Impact of Aggregation on Standard Error for Smaller Geographies, example of the Percentage of Children under Age 18 in Poverty

An estimate of child poverty can be obtained from ACS table DP03 as well as from ACS tables B17020 and B17001. Similarly to the cases in the previous example, the advantage of using DP03 over a B-series table is that the standard error for the estimate does not have to be approximated, as it is already provided by the Census Bureau. The disadvantage of the DP03 table is that estimates cannot be obtained for population subgroups (except when using the Selected Population Tables version, which is published every 5 years). On the other hand, B-series tables provide data for population subgroups (sex, race/ethnicity) and data extraction can be automated via the API. However, the standard error has to be approximated by aggregating two or more categories (e.g., summing over multiple poverty levels).

This example presents a comparison of pre-calculated (from table DP03) and approximated standard errors (from table B17020) using child poverty in California data for 2010–2014 to show how approximation can over- or under-estimate standard errors. The results are shown in Table 3. The pre-calculated and the approximated standard error distributions differ and are affected by the geographic level. County distributions are very similar, but the place and census tract distributions for approximated standard errors have much larger dispersion and maximum values than the pre-calculated standard errors.

The association between DP03 standard error and approximated B17020 standard error is shown in Figure 2a. The association is more or less linear. The maximum value for DP03 standard errors was 60.79 for census tracts and places but can be as high as 1,938 for B17020. Figure 2b shows the association when all cases in which the DP03 standard error is equal to 60.79 are excluded; this figure illustrates that approximated standard errors could be over or under estimated.

Table 4 shows which geographies in California have the highest DP03 standard errors and approximated B17020 standard errors. Table B17020 provide the numerator and denominator for the estimate and it is possible to see that high errors occur when low counts are observed.

In conclusion, this examination shows that there are limitations to using B-series tables to obtain standard errors of estimates. The use of these standard errors to produce confidence intervals or conduct statistical tests should be done with caution.

Table 3. Comparison of the Distribution of the Standard Error from Table DP03 and the Distribution of Approximated Standard Error for Table B17020 for Child Poverty Indicator. California, 2010-2014.

Distribution	B17020_StdErr	DP03_StdErr
Census Tract		
Min	0.46	0.12
1st Quartile	3.64	4.40
Median	5.58	7.11
Mean	6.62	7.39
3rd Quartile	7.45	9.30
Max	499.40	60.79
NA's	98	98
Place		
Min	0.24	0.24
1st Quartile	2.45	2.98
Median	5.38	6.81
Mean	16.81	11.42
3rd Quartile	12.23	15.02
Max	1938.00	60.79
NA's	116	116
County		
Min	0.14	0.18
1st Quartile	0.57	0.74
Median	1.06	1.37
Mean	1.53	1.91
3rd Quartile	2.15	2.72
Max	7.86	8.02
NA's		

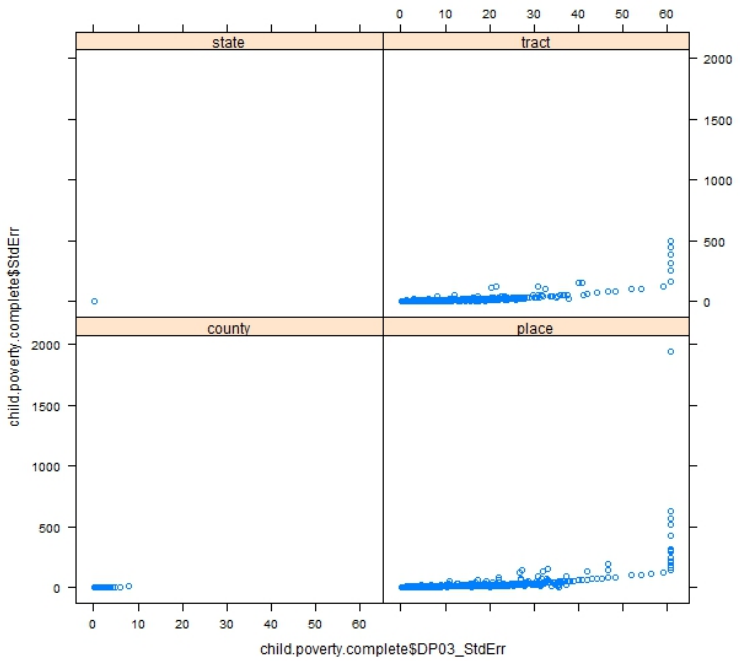


Figure 3a. XY plot showing the relationship between DP03 standard error (x) and approximated B17020 standard error (y), for state, tract, county and place geographies for California, 2010-2014.

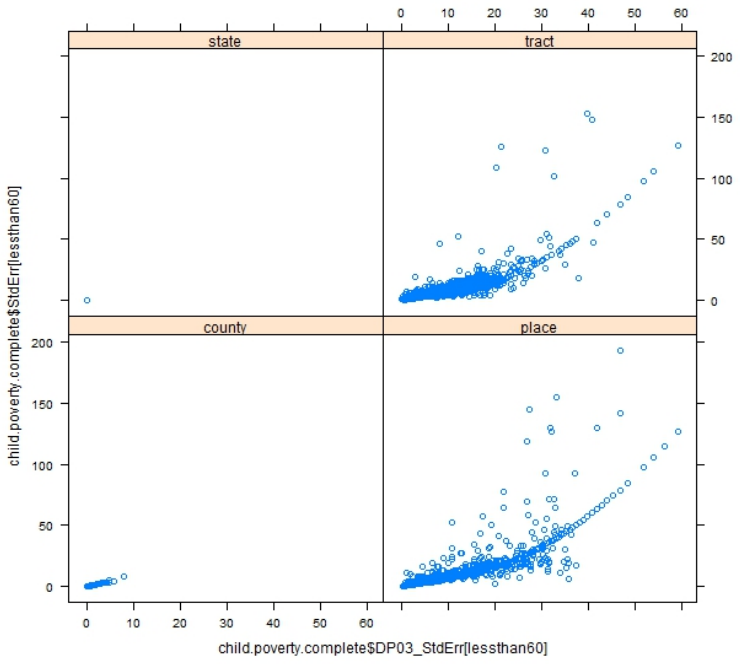


Figure 3b. XY plot showing the relationship between DP03 standard error (x) and approximated B17020 standard error (y), for state, tract, county and place geographies; the plot excludes cases when DP03 standard error is higher than 60.79. California, 2010-2014.

Table 5. California Geographies with the Largest Standard Error Terms for Table DP03 and Table B71020, 2010-2014.

NAME	Percent DP03	StdErr	Numerator B71020	Denominator	Percent	StdErr B71020
Alpine Village CDP, California	0	60.79	0	7	0	180.50
Big Bend CDP, California	0	60.79	0	9	0	140.39
Bridgeport CDP, California	0	60.79	0	6	0	210.58
California Hot Springs CDP, California	0	60.79	0	8	0	157.94
Edgewood CDP, California	0	60.79	0	7	0	180.50
Elk Creek CDP, California	100	60.79	4	4	100	512.23
Fields Landing CDP, California	100	60.79	8	8	100	290.15
Furnace Creek CDP, California	100	60.79	1	1	100	1937.6
Gazelle CDP, California	100	60.79	3	3	100	567.38
Hat Creek CDP, California	100	60.79	9	9	100	247.81
Idlewild CDP, California	0	60.79	0	4	0	315.88
Lake Almanor West CDP, California	0	60.79	0	6	0	210.58
Lodoga CDP, California	0	60.79	0	6	0	210.58
Moss Landing CDP, California	100	60.79	7	7	100	307.04
Paynes Creek CDP, California	0	60.79	0	3	0	421.17
Redcrest CDP, California	0	60.79	0	4	0	315.88
Timber Cove CDP, California	0	60.79	0	2	0	631.75
Census Tract 5.04, Calaveras County, California	0	60.79	0	4	0	315.88
Census Tract 4032, Los Angeles County, California	100	60.79	5	5	100	446.05
Census Tract 5755, Los Angeles County, California	0	60.79	0	8	0	157.94
Census Tract 109, Monterey County, California	0	60.79	0	8	0	250.07
Census Tract 9800, Monterey County, California	100	60.79	5	5	100	389.44
Census Tract 9883, Sacramento County,	0	60.79	0	4	0	447.49
Census Tract 57, San Diego County, California	0	60.79	0	4	0	315.88
Census Tract 9802, San Francisco County,	0	60.79	0	8	0	157.94
Census Tract 5116.08, Santa Clara County,	100	60.79	4	4	100	499.44
Census Tract 1516.01, Sonoma County, California	0	60.79	0	8	0	157.94

Measure development using Federal Bureau of Investigations Uniformed Crime Reports: Number of Reported Violent Crime Offenses per 100,000 Population

The Federal Bureau of Investigation publishes annual Uniform Crime Reports on its website, including Excel tables and the newer Data Tool (https://www.ucrdatatool.gov/).

As highlighted before, the focus of this project was in the automation of the production of datasets. At the time of the project (Fall 2015), one of the challenges with the publicly available UCR Excel tables was their formatting making them not machine readable ready.

The Federal Bureau of Investigation was contacted to obtain copies of the Uniformed Crime Reports "Crime by County" files produced for California between the years 2000 and 2013 (https://www.fbi.gov/about-us/cjis/ucr/ucr). An email was sent to CRIMESTATSINFO@ic.fbi.gov on September 28th 2015, and a positive response was received on October 20th 2015. The files received were text files that include data for all law enforcement agencies in the United States that participate in the UCR program. The formatting in these files allowed for faster import into statistical software for data extraction.

The UCR text files provide information by county and by agency within the county. Some of the agencies correspond to cities or towns, some others correspond to Sheriff offices, tribal law enforcement, university law enforcement, Bay Area Rapid Transit law enforcement, among others. The total crimes for cities and towns in a county are summed with the crimes reported by all other agencies in that county to produce a county total. The UCR text files do not provide U.S. Census geographical codes for cities and towns or counties. They do provide agency codes (ORI7) that can be merged with a crosswalk table that bridges agency codes to U.S. Census geographical codes. The crosswalk table was obtained from the following source: *United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. Law Enforcement Agency Identifiers Crosswalk, 2012. ICPSR35158-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-04-17. http://doi.org/10.3886/ICPSR35158.v1*. After exploration it was determined that the crosswalk table did not include 3 law reporting agencies, which were added manually. The table also included 4 agency duplicates that were removed.

Currently it is not possible to obtain crime data for geographical areas below the city level. Additionally, it is not possible to obtain specific crime data for jurisdictions that do not have their own law enforcement agency.

Results

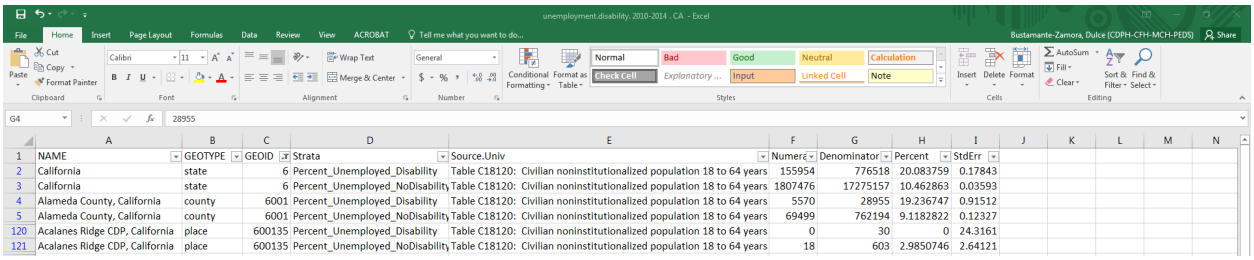
The deliverables for the pilot project consisted of Excel or csv files with the data for California, its counties, cities and towns, and census tracts. For the AskCHIS NE data, geographies included the state, counties, cities and towns, and ZCTAs. For the crime data, only county and city data were provided. For some measures, data for New York State was provided in order to demonstrate that automation of dataset production was feasible. Additionally, R and SAS code was generated for the automation of the construction or calculation of the measures. The complete list can be found in Table 6.

Table 6. Complete List of Final Version Files for each of the Indicators Produced During the California Pilot

File name	Content
Percentage of children under age 18 in poverty (race/ethnicity and sex strata)	
child.poverty.raceeth_acs.B17020.R	R file with code for extracting data from table B17020
child.poverty.sex_acs.B17001.R	R file with code for extracting data from table B17001
child.poverty.sex. 2010-2014 . CA.xlsx, child.poverty.sex. 2010-2014 . NY.xlsx	Output files containing indicator estimate and its standard error with sex strata for California and New York (census tract, place, county, state)
child.poverty.race. 2010-2014 . NY.R, child.poverty.race. 2010-2014 . CA.R	Output files containing indicator estimate and its standard error with race/ethnicity strata for California and New York (census tract, place, county, state)
Ratio of household income at the 80th percentile to income at the 20th percentile	
incomeInequality.acs.B19080.R	R file with code for extracting data from table B19080
income.inequality. 2010-2014 . NY.xlsx, income.inequality. 2010-2014 . CA.xlsx	Output files containing indicator estimate and its standard error for California and New York (census tract, place, county, state)
Percentage of adults ages 25-44 years with some post-secondary education (race/ethnicity and sex strata)	
some.college.raceeth_acs.B15002.R	R file with code for extracting data from table B15002
some.college.sex_acs.B15001.R	R file with code for extracting data from table B15001
some.college.race. 2010-2014 . CA.xlsx, some.college.race. 2010-2014 . NY.xlsx	Output files containing indicator estimate and its standard error with race/ethnicity strata for California and New York (census tract, place, county, state)
some.college.sex. 2010-2014 . CA.xlsx, some.college.sex. 2010-2014 . NY.xlsx	Output file containing indicator estimate and its standard error with sex strata for California and New York (census tract, place, county, state)
Percentage of the workforce that drives alone to work (race/ethnicity, sex, and poverty level strata)	
drove.alone.sex_acs.B08006.R	R file with code for extracting data from table B08006
drove.alone.sex. 2010-2014 . NY.xlsx, drove.alone.sex. 2010-2014 . CA.xlsx	Output file containing indicator estimate and its standard error with sex strata for California and New York (census tract, place, county, state)
drove.alone.raceeth_acs.B08105.R	R file with code for extracting data from table B08105
drove.alone.race. 2010-2014 . NY.xlsx, drove.alone.race. 2010-2014 . CA.xlsx	Output files containing indicator estimate and its standard error with race/ethnicity strata for California and New York (census tract, place, county, state)
drove.alone.poverty_acs.B08122.R	R file with code for extracting data from table B08122
drove.alone.poverty. 2010-2014 . NY.xlsx, drove.alone.poverty. 2010-2014 . CA.xlsx	Output files containing indicator estimate and its standard error with poverty strata for California and New York (census tract, place, county, state)

Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing (race/ethnicity strata)	
severe.housing.problems.chas.2009.2013.R	R file with code for extracting data from Table 2 from HUD-CHAS 2009-013 data set
severe.housing.problems.ct.2009-2013.csv, severe.housing.problems.co.2009-2013.csv, severe.housing.problems.pl.2009-2013.csv, severe.housing.problems.st.2009-2013.csv	Output files containing indicator estimate and its standard error with race/ethnicity strata for all States in the country. One file per geographic level: census tract (ct), place (pl), county (co), and state (st)
Percentage of adults reporting fair or poor health **Percentage of adults that report a BMI of 30 or more (race/ethnicity, sex strata)** **Percentage of population under age 65 without health insurance**	
CHIS.fph.CA.2014.csv, CHIS. obese.CA.2014.csv, CHIS.unins.CA.2014.csv, CHIS.obese.race.sex.CA.2014	Output files containing indicator estimate and its standard error for all three indicators indicated above. There is one file per geographical level: ZCTA, counties and cities. CHIS.obese.race.sex.CA.2014 contains BMI modeled data by race/ethnicity and gender that was produced for this project
Percentage of population ages 16 and older unemployed but seeking work (race/ethnicity, sex, poverty, and disability strata)	
unemployment.disability_acs.C18120	R file with code for extracting data from table C18120
unemployment.sex.poverty_acs.B17005	R file with code for extracting data from table B17005
unemployment.raceeth_acs.C23002	R file with code for extracting data from table C23002
unemployment.disability. 2010-2014 . NY. csv, unemployment.disability. 2010-2014 . CA.csv	Output files containing indicator estimate and its standard error with disability strata for California and New York (census tract, place, county, state)
unemployment.race. 2010-2014 . CA.csv, unemployment.race. 2010-2014 . NY.csv	Output files containing indicator estimate and its standard error with race/ethnicity strata for California and New York (census tract, place, county, state)
unemployment.sex. 2010-2014 . NY. csv, unemployment.sex. 2010-2014 . CA. csv	Output files containing indicator estimate and its standard error with sex strata for California and New York (census tract, place, county, state)
unemployment.poverty. 2010-2014 . CA.csv, unemployment.poverty. 2010-2014 . NY.csv	Output files containing indicator estimate and its standard error with poverty strata for California and New York (census tract, place, county, state)
Number of reported violent crime offenses per 100,000 population	
HCI_Crime_Wisconsin-with coverage-1-4-16.sas	R file with code for extracting data from the UCR FBI Crime by County text files (UCR 55100) for California, including the coverage adjustment following County Health Rankings and Roadmaps methodoloy
HCI_Crime_Wisconsin_PL_CO_RE_CA_20052007_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20062008_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20072009_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20082010_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20092011_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20102012_06JAN16.xlsx, HCI_Crime_Wisconsin_PL_CO_RE_CA_20112013_06JAN16.xlsx	Output files containing indicator estimate and its standard error for California

Output files contained as a minimum nine fields of information: locality name, geotype (type of geographic unit), geoid (geographic code), strata, source and universe, numerator, denominator, estimate and standard error. A screenshot of the output file can be found below.



	A	B	C	D	E	F	G	H	I
1	NAME	GEOTYPE	GEOID	Strata	Source.Univ	Numera	Denominator	Percent	StdErr
2	California	state	6	Percent_Unemployed_Disability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	155954	776518	20.083759	0.17843
3	California	state	6	Percent_Unemployed_NoDisability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	1807476	17275157	10.462863	0.03593
4	Alameda County, California	county	6001	Percent_Unemployed_Disability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	5570	28955	19.236747	0.91512
5	Alameda County, California	county	6001	Percent_Unemployed_NoDisability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	69499	762194	9.1182822	0.12327
120	Acalanes Ridge CDP, California	place	600135	Percent_Unemployed_Disability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	0	30	0	24.3161
121	Acalanes Ridge CDP, California	place	600135	Percent_Unemployed_NoDisability	Table C18120: Civilian noninstitutionalized population 18 to 64 years	18	603	2.9850746	2.64121

The datasets that were produced during the CHR&R pilot project were not published in the HCI website due to competing priorities. However, the measures developed for this pilot project that did not overlap with the HCI measures have been incorporated into the HCI project. The HCI disseminates data via its website: https://www.cdph.ca.gov/Programs/OHE/Pages/HCI-Search.aspx#ALoESD. The only dataset from the pilot project that was published online was the crime rate dataset: https://www.cdph.ca.gov/Programs/OHE/CDPH%20Document%20Library/HCI/ADA%20Compliant%20Documents/HCI_Crime_752_PL_CO_RE_CA_2000-2013_21OCT15-ADA.xlsx. As with all HCI measures, a companion narrative file with metadata was also published: https://www.cdph.ca.gov/Programs/OHE/CDPH%20Document%20Library/HCI/ADA%20Compliant%20Documents/HCI_Crime_752-Narrative_Examples-10-30-15-ADA.pdf.

The HCI does not currently have an evaluation plan to determine usage and reach. However, there are multiple use cases for the project methods and data in local health departments. One relevant example is a guide for local health departments that was developed by the Bay Area Regional Inequities Initiative, Data Committee: http://www.barhii.org/download/publications/barhii_sdoh_indicator_guide_v1.1.pdf, that was partially informed by the methods of the HCI project.

Project Sustainability

The HCI project is committed to provide disaggregated data for California communities that can help assess community health and equity, and can inform the planning of healthier communities. The HCI project exists within the Office of Health Equity, which has a legislative mandate in California to provide information to the people of state on the "underlying conditions that contribute to health and well-being."

The HCI received funding from the California Strategic Growth Council (SGC) from 2012-2014 and from this CHR&R Pilot Project between 2015 and 2016. During the SGC funded period the HCI project had 3 full time equivalent (FTE) employees. During the CHR&R funded period the project had only 1 FTE. To ensure the project's sustainability the OHE has partnered internally with other indicator projects in the California Department of Public Health to join forces in delivering data for public health reporting and analysis. These partnerships have received the support of CDPH leadership and in 2015 OHE received approval to hire one new staff position dedicated to the HCI. A new staff member came onboard in February 2017, increasing our staff to 2 FTEs. Additionally, CDPH now provides programs new data visualization tools (at no extra cost). HCI is developing data visualizations with Tableau software and ESRI Story Maps that make the data more accessible to a general audience.

Technical Lessons Learned

Challenges to Studying Changes over Time for Sub-county Geographies

The changing nature of sub-county geographies is a challenge for creating standard measures that are comparable over time. Census designated places and census tracts can go through mergers or dissolutions during censal and intercensal periods. We have created a lookup table that shows these changes over time (2000 to present) for California but similar tables would need to be created for each state, and they would need to be continuously updated. For measures that are extracted directly from the U.S. Census, there are private companies and universities that offer algorithms to bridge geographies and help recalculate measures to reflect the mergers or dissolutions, at least at the Census tract level. One example is the Longitudinal Tract Data Base from Brown University.

ANNEX I

R Code for Automated Data Download and Generation of Unemployment by Disability Status Dataset

```
#Dulce  Bustamante, 6-1-16
#Office of Health Equity
#California Department of Public Health

#########################################################################
#Extraction of ACS data on unemployment by disability status
#by state, county, place, and census tract.

#ACS table C18120 is "fetched" from the Census Application Programming
#Interface (http://www.census.gov/developers/) using the acs.R package (Glenn, 2011).

#User should specify desired year, span, state, time range
# ACS data set options available (as of 5-12-16):
# 1 year files: 2014, 2013, 2012, 2011 (if this is selected an error message will
appear since tracts are not available)
# 3 year files: 2011-2013, 2010-2012
# 5 year files: 2010-2014, 2009-2013, 2008-2012, 2007-2011, 2006-2010, 2005-2009

#Table C18120 contains "EMPLOYMENT STATUS BY DISABILITY STATUS"
#The Universe:  Civilian noninstitutionalized population 18 to 64 years

#The data is "fetched" and used to calculate, for each geography,
#(1) Percent_Unemployed_Disability: percentage of individuals in the labor force (18-
64 years) and with a disability that are unemployed and standard error
#(2) Percent_Unemployed_NoDisability: percentage of individuals in the labor force
(18-64 years) and without a disability that are unemployed and standard error

#Before extracting data using the acs.R, it is necessary to obtain a Key at the
#the Census API developers site.  Enter key at line 40.
#########################################################################

#########################################################################
#Loading packages
#########################################################################
library(acs)
library(reshape)

#########################################################################
#Before extracting data using the acs.R, it is necessary to obtain a Key at the
#the Census API developers site; it takes two minutes to obtain a key.
#########################################################################
#My API key installation
#api.key.install(key="...enter your key here... ")

#########################################################################
#Example of acs.lookup to search for keyword matches in the ACS metadata
#########################################################################
#disability.unemployed <- acs.lookup(endyear=2014, span=5, table.name=c("Disability"))

#########################################################################
#Set working directory
#########################################################################
setwd("T:\\HCI\\BusinessPlan\\CountyHealthRankings\\acs.RpackageNYcode\\Unemployment\\
")

#########################################################################
#Specify here the end year (), the span (5, 3 or 1 year file), the
```

```
#state, and the year range for the data extraction
#Note: do not modify other lines of code
#########################################################################
year<-2014
span<-5
yearrange<-c("2010-2014")
state_id<-"CA"

#########################################################################
#Creating geographies to extract data for the state, all counties, places and tracts
#########################################################################
state  <- geo.make(state = state_id)
county <- geo.make(state = state_id, county="*")
place  <- geo.make(state = state_id, place="*")
tract  <- geo.make(state = state_id, county="*",tract="*")
#geo.multiple<-geo.state+geo.county+geo.city+geo.tract

geolist<-c(state,county,place,tract)
geonames<-c("state", "county","place","tract")


#########################################################################
#Creating empty data frames to add formated data
#########################################################################

unemployment.disability.total  <- data.frame("NAME"=NA, "GEOTYPE"=NA, "GEOID"=NA,
"Strata"=NA,
                                        "Numerator"=NA, "Denominator"=NA, "Percent"=NA,
"StdErr"=NA)


#########################################################################
#Loop for data extraction and formating
#########################################################################

for (i in 1:length(geolist)) {

  #"Fetching" data from the Census API to create acs-class object
  unemployment<-acs.fetch(endyear=year, span=span, geography = geolist[i],
table.number = "C18120", dataset="acs", col.names="pretty")

  #Using the divide.acs function to calculate proportions
  #The divide function also calculates the standard error following the Census
specifications

  #Percent of unemployment, with a disability
  unemployment.disability   <-
divide.acs(unemployment[,7],unemployment[,4]+unemployment[,7],method="proportion")
  #Explanation of variables
  #unemployment[,7]  = Employment Status by Disability Status: In the labor force:
Unemployed: With a disability
  #unemployment[,4]  = Employment Status by Disability Status: In the labor force:
Employed: With a disability


  #Percent of unemployment, without a disability
  unemployment.nodisability  <-
divide.acs(unemployment[,8],unemployment[,8]+unemployment[,5],method="proportion")
  #unemployment[,8]  = Employment Status by Disability Status: In the labor force:
Unemployed: No disability
  #unemployment[,5] = Employment Status by Disability Status: In the labor force:
Employed: No disability
```

```
  #Calculating numerator and denominator to add to table
  unemployment.numerator.disability   <-unemployment[1:dim(unemployment)[1],7]
  unemployment.denominator.disability <-
unemployment[1:dim(unemployment)[1],4]+unemployment[1:dim(unemployment)[1],7]

  unemployment.numerator.nodisability   <-unemployment[1:dim(unemployment)[1],8]
  unemployment.denominator.nodisability <-
unemployment[1:dim(unemployment)[1],8]+unemployment[1:dim(unemployment)[1],5]

  #Add meaningful names for the percent estimate, standard error, numerator and
denominator
  #These names will be later used to create strata names

  dimnames(unemployment.disability@estimate)[[2]]       <-
paste("Est_Percent_Unemployed_Disability")
  dimnames(unemployment.disability@standard.error)[[2]] <-
paste("StE_Percent_Unemployed_Disability")

  dimnames(unemployment.nodisability@estimate)[[2]]       <-
paste("Est_Percent_Unemployed_NoDisability")
  dimnames(unemployment.nodisability@standard.error)[[2]] <-
paste("StE_Percent_Unemployed_NoDisability")

  dimnames(unemployment.numerator.disability@estimate)[[2]]   <-
paste("Num_Percent_Unemployed_Disability")
  dimnames(unemployment.denominator.disability@estimate)[[2]] <-
paste("Den_Percent_Unemployed_Disability")

  dimnames(unemployment.numerator.nodisability@estimate)[[2]]   <-
paste("Num_Percent_Unemployed_NoDisability")
  dimnames(unemployment.denominator.nodisability@estimate)[[2]] <-
paste("Den_Percent_Unemployed_NoDisability")


  #Create a data frame for disability data  with the acs object data
  unemployment.disability_df  <-
data.frame("NAME"=unemployment.disability@geography$NAME,
                                          "GEOTYPE"=geonames[i],

"GEOID"=paste0(str_pad(unemployment.disability@geography$state, 2, "left", pad="0"),

str_pad(unemployment.disability@geography$county,3, "left", pad="0"),

str_pad(unemployment.disability@geography$place, 5, "left", pad="0"),

str_pad(unemployment.disability@geography$tract, 6, "left", pad="0")),

unemployment.numerator.disability@estimate,
unemployment.numerator.nodisability@estimate,

unemployment.denominator.disability@estimate,
unemployment.denominator.nodisability@estimate,
                                          unemployment.disability@estimate,
unemployment.disability@standard.error,
                                          unemployment.nodisability@estimate,
unemployment.nodisability@standard.error,
                                          stringsAsFactors=FALSE)

  #Creating separate files for numerator, denominator, percent estimate and standard
error
  unemployment.disability.Num_df<-unemployment.disability_df[ ,c(1,2,3,4,5)]
  unemployment.disability.Den_df<-unemployment.disability_df[ ,c(1,2,3,6,7)]
```

```
  unemployment.disability.Est_df<-unemployment.disability_df[ ,c(1,2,3,8,10)]
  unemployment.disability.SE_df<-unemployment.disability_df[ ,c(1,2,3,9,11)]

  #Transposing the numerator, denominator, percent estimate and standard error files
  unemployment.disability.Num.transp_df <- melt(unemployment.disability.Num_df,
id=c("NAME","GEOTYPE","GEOID"), variable_name="Strata")
  unemployment.disability.Den.transp_df <- melt(unemployment.disability.Den_df,
id=c("NAME","GEOTYPE","GEOID"), variable_name="Strata")
  unemployment.disability.Est.transp_df <- melt(unemployment.disability.Est_df,
id=c("NAME","GEOTYPE","GEOID"), variable_name="Strata")
  unemployment.disability.SE.transp_df <- melt(unemployment.disability.SE_df,
id=c("NAME","GEOTYPE","GEOID"), variable_name="Strata")

  #Renaming numerator, denominator, percent estimate and standard error columns

names(unemployment.disability.Num.transp_df)[names(unemployment.disability.Num.transp_
df)=="value"] <- "Numerator"

names(unemployment.disability.Den.transp_df)[names(unemployment.disability.Den.transp_
df)=="value"] <- "Denominator"

names(unemployment.disability.Est.transp_df)[names(unemployment.disability.Est.transp_
df)=="value"] <- "Percent"

names(unemployment.disability.SE.transp_df)[names(unemployment.disability.SE.transp_df
)=="value"] <- "StdErr"

  #Updating strata name to character type
  unemployment.disability.Num.transp_df$Strata<-
as.character(unemployment.disability.Num.transp_df$Strata)
  unemployment.disability.Den.transp_df$Strata<-
as.character(unemployment.disability.Den.transp_df$Strata)
  unemployment.disability.Est.transp_df$Strata<-
as.character(unemployment.disability.Est.transp_df$Strata)
  unemployment.disability.SE.transp_df$Strata<-
as.character(unemployment.disability.SE.transp_df$Strata)

  #Removing the first 4 characters from names to create strata names
  for (j in 1:dim(unemployment.disability.Num.transp_df)[1])
    unemployment.disability.Num.transp_df$Strata[j]<-
substr(unemployment.disability.Num.transp_df$Strata[j],5,nchar(unemployment.disability
.Num.transp_df$Strata[j]))

  for (j in 1:dim(unemployment.disability.Den.transp_df)[1])
    unemployment.disability.Den.transp_df$Strata[j]<-
substr(unemployment.disability.Den.transp_df$Strata[j],5,nchar(unemployment.disability
.Den.transp_df$Strata[j]))

  for (j in 1:dim(unemployment.disability.Est.transp_df)[1])
    unemployment.disability.Est.transp_df$Strata[j]<-
substr(unemployment.disability.Est.transp_df$Strata[j],5,nchar(unemployment.disability
.Est.transp_df$Strata[j]))

  for (j in 1:dim(unemployment.disability.SE.transp_df)[1])
    unemployment.disability.SE.transp_df$Strata[j]<-
substr(unemployment.disability.SE.transp_df$Strata[j],5,nchar(unemployment.disability.
SE.transp_df$Strata[j]))

  #Merge files
  unemployment.disability1.data <- merge(unemployment.disability.Num.transp_df,
unemployment.disability.Den.transp_df, by=c("NAME", "GEOTYPE", "GEOID", "Strata"))
  unemployment.disability2.data <- merge(unemployment.disability.Est.transp_df,
unemployment.disability.SE.transp_df, by=c("NAME", "GEOTYPE", "GEOID", "Strata"))
```

```
  unemployment.disability.data <- merge(unemployment.disability1.data,
unemployment.disability2.data, by=c("NAME", "GEOTYPE", "GEOID", "Strata"))

  #Convert proportions to percentage
  unemployment.disability.data$Percent <- unemployment.disability.data$Percent*100
  unemployment.disability.data$StdErr <- unemployment.disability.data$StdErr*100

  #unemployment.disability.data<-sort_df(unemployment.disability.data,vars =
c("GEOTYPE", "GEOID", "Strata"))

  #Bind disability data to total file
  unemployment.disability.total<-
rbind(unemployment.disability.total,unemployment.disability.data)

}

#############################################################################
#Final formatting and data export
#############################################################################

unemployment.disability.total<-unemployment.disability.total[-1,]

source.univ.col <- rep("Table C18120:  Civilian noninstitutionalized population 18 to
64 years",dim(unemployment.disability.total)[1])

unemployment.disability.total$Source.Univ<-source.univ.col

unemployment.disability.total<-unemployment.disability.total[,c(1,2,3,4,9,5,6,7,8)]

filename <- paste("unemployment.disability.",yearrange[1],".",state_id,".csv")

write.csv(unemployment.disability.total, file=filename,row.names=FALSE)
```

ANNEX II

Example Calculations for the Estimated Unemployment Rate, its Numerator, Denominator and Standard Error as it appears in Table 2.

Census Product	Universe	Topic	Comments	Example 2009-2013 Data, California
S2301: Employment Status	Population 16 years and over	Disability	Includes total number of adults 16 years and over with any disability and the percent of those in the labor force and the percent of those in the labor force but unemployed. The table does not present the estimated count of adults with any disability that are in the labor force or unemployed; these counts need to be calculated manually.	Population 16 years an over with any disability: 1,830,156 +/- 9,340 Margin of error Population 16 years an over in the labor force 41.1% +/- 0.3 Population 16 years an over in the labor force - Unemployed (Unemployment estimate) 20.0% +/- 0.4 (Margin of error) SE=0.4/1.645=0.24, based on the formula Standard Error = Margin of Error / 1.645 <u>Example manual calculation of estimates:</u> Population 16 years an over with any disability in the labor force (denominator): 1,830,156*0.411=752,194 Population 16 years an over with any disability in the labor force-Unemployed (numerator): 1,830,156*0.411*0.2=150,439
C18120: Employment Status by Disability Status	Civilian noninstitutionalized population 18-64 years	Disability	Table presents estimated counts of adults and the margin of error of the count, but no percent estimates. Universe does not include adults over 65 years of age.	People with a disability in the labor force – Employed 609,443 +/- 6,218 SE=3,780 People with a disability in the labor force – Unemployed (numerator) 157,588 +/- 3,282 SE (numerator)=1,995 <u>Example manual calculation of percentages:</u> People with a disability in the labor force (denominator): 609,443 + 157,588 = 767,031 SE=sqrt($3{,}780^2$+$1{,}995^2$)=4,274, calculated using the formula: $SE(\hat{X}_1 \pm \hat{X}_2) \approx \sqrt{(SE(\hat{X}_1))^2 + (SE(\hat{X}_2))^2}$ Percent of people with a disability in the labor force – Unemployed (unemployment estimate): 157,588 / 767,031 * 100 = 20.5% SE=0.23, calculated using approximate method formula below

				$SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$
B23024: Poverty Status in the Past 12 Months by Disability Status by Employment Status for the Population 20 to 64 years	Population 20 to 64 years for whom poverty status is determined	Disability Poverty	Table presents estimated number of adults and the margin of error, but no percent estimates. Universe does not include adults less than 18 years of age and over 65 years of age.	Below poverty, with a disability, civilian, Employed 66,342 +/- 1,439 SE=875 Below poverty, with a disability, civilian, Unemployed 56,045 +/- 1,941 SE=1,180 Above poverty, with a disability, civilian, Employed 532,980 +/- 5,888 SE=3,579 Above poverty, with a disability, civilian, Unemployed 93,754 +/- 2,465 SE=1,498 <u>Example manual calculation of percentages:</u> People with a disability, civilian (denominator): 66,342+56,045+532,980+93,754=749,121 SE=sqrt(875^2+$1,180^2$+$3,579^2$+$1,498^2$)=4,149 People with a disability, civilian unemployed (numerator): 56,045+93,754=149,799 SE=sqrt($1,180^2$+$1,498^2$)=1,907 Percent people with a disability, civilian unemployed (Unemployment): 149,799 / 749,121 * 100 = 20.0% SE=0.23, calculated using approximate method formula below $SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$
B23001: Sex by Age by Employment Status for the Population 16 years and Over	Population 16 years and over	Sex	The table presents employment status by sex and broken down by 13 different age categories. 13 different estimates would need to be used to approximate a total male or female estimate and standard error. This	Note: there are 32 data points for the males in the labor force and those unemployed. <u>Results manual calculation male unemployment rate:</u> Civilian Males in Labor Force: 9,869,815 SE=162,525 Civilian Males in Labor Force -Unemployed: 1,188,964

			is a large number of estimates and the standard error obtained would be very large or very small.	SE= 4,651 Percent unemployed males: 12.0% SE=0.20
S2301: Employment Status	Population 16 years and over Subset population 20 to 64 years	Sex	Includes total number of adults 20 to 64 years and the percent of those in the labor force and those in the labor force but unemployed. The table does not present the estimated number of adults by sexs that are in the labor force or unemployed; these counts need to be calculated manually.	Population 20 to 64 years - MALE: 11,474,753 +/- 953 In the labor force 83.0% +/- 0.1 Unemployed (Unemployment) 10.9% +/- 0.1 SE=0.06 <u>Example manual calculation of estimates:</u> Population 20 to 64 years - MALE in the labor force: 11,474,753*0.83=9,524,045 Population 20 to 64 years - MALE in the labor force-Unemployed: 11,474,753*0.83*0.109= 1,038,121
B17005: Poverty Status in the Past 12 Months of Individual by Sex by Employment Status	Civilian population 16 years and over for whom poverty status is determined	Sex	This table excludes people 16 years and over for whom poverty status has not been determined (~2.3%)	Income Below Poverty Level – Male – In Labor Force: 962,972 +/- 8,318 SE=5,056 Income Below Poverty Level – Male – In Labor Force - Unemployed: 315,842 +/- 4,944 SE=3,005 Income Above Poverty Level – Male – In Labor Force: 9,174,641 +/- 12,631 SE=7,678 Income Above Poverty Level – Male – In Labor Force - Unemployed: 867,603 +/- 7,149 SE=4,345 <u>Example manual calculation male unemployment rate:</u> Males in Labor Force (denominator): 962,972+9,174,641=10,137,613 SE=sqrt($5,056^2$+$7,678^2$)=9,193 Males in Labor Force –Unemployed (numerator): 315,842+867,603=1,183,445

				SE=sqrt($3{,}005^2$+$4{,}345^2$)=5,284 Percent unemployed males: 1,183,445/10,137,613*100=11.7% SE=0.05, calculated using approximate method formula below $SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$
S2301: Employment Status	Population 16 years and over	Poverty	Includes total number of adults 16 years and over below poverty and the percent of those in the labor force and those in the labor force but unemployed. The table does not present the estimated number of adults below poverty that are in the labor force or unemployed; these counts need to be calculated manually.	Population below poverty level: 3,208,070 +/- 20,819 In the labor force 53.4% +/- 0.2 Unemployed (Unemployment) 31.7% +/- 0.3 SE=0.18 <u>Example manual calculation of estimates:</u> Population below poverty level in the labor force: 3,208,070*0.534=1,713,109 Population below poverty level in the labor force-Unemployed: 3,208,070*0.534*0.317=543,056
B17005: Poverty Status in the Past 12 Months of Individual by Sex by Employment Status	Civilian population 16 years and over for whom poverty status is determined	Poverty	This table excludes people 16 years and over for whom poverty status has not been determined (~2.3%)	Income Below Poverty Level – Male – In Labor Force: 962,972 +/- 8,318 SE=5,056 Income Below Poverty Level – Male – In Labor Force - Unemployed: 315,842 +/- 4,944 SE=3,005 Income Below Poverty Level – Female – In Labor Force: 896,339 +/- 7,544 SE=4,586 Income Below Poverty Level – Female – In Labor Force - Unemployed: 293,996 +/- 4,301 SE=2,614 <u>Example manual calculation male unemployment rate:</u> People below poverty level in Labor Force (denominator): 962,972+896,339=1,859,311

				SE=sqrt($5{,}056^2$+$4{,}586^2$)=6,826 People below poverty level in Labor Force –Unemployed (numerator): 315,842+293,996=609,838 SE=sqrt($3{,}005^2$+$2{,}614^2$)=3,984 Percent unemployed below poverty level: 609,838/1,859,311=32.8% SE=0.18, calculated using the approximate method formula below $SE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[SE(\hat{X})]^2 - \frac{\hat{X}^2}{\hat{Y}^2}[SE(\hat{Y})]^2}$

New York State

Introduction

This white paper reports a New York State research team's methods and results from a 2016 sub-county health data pilot project. It also provides supplemental materials that (e.g., datasets, reference files, SAS code) that were used for sub-county data reporting. The authors would like this paper to serve as a useful resource to support other researchers in carrying out similar projects in the future.

Aims of the pilot project

1. To identify measures from the County Health Rankings and Roadmaps model that are appropriate for sub-county level analysis.
2. To generate the identified measures at sub-county levels.
3. To assess the impact of data suppression and estimate instability on the generated estimates.
4. To disseminate the data to public health practitioners in the form of county-specific health indicator reports, designed to facilitate targeted interventions for community health improvement.

Intended Audiences

The research team aimed to provide small area data to analysts and public health practitioners, including those among local health departments, hospitals, regional health planning organizations, and community-based organizations. The research team hoped to positively affect the health of disparate communities and populations in New York State by providing the data.

Methods

Environmental Scan of Data Sources

New York's original project proposal was to generate sub-county data for measures that aligned with the county health rankings model. Therefore, the environmental scan of data sources involved determining which sub-county data sources could possibly be used to re-create County Health Rankings measures with fidelity, and among them, what years of data were available. The research team examined definitions for each of the County Health Rankings measures, and then identified the data sources that would likely be able to provide corresponding sub-county data. The research team set out to precisely follow the definitions for most of the County Health Rankings measures; however, after conducting the first environmental scan of the data sources and the core County Health Rankings measures, the team decided to modify its

definitions for certain measures when data for the exact indicator were not available. Three data sources were selected: New York State Vital Records (birth and death data), New York State SPARCS (hospitalization data), and the New York State 2013-2014 Expanded Behavioral Risk Factor Surveillance Survey (health prevalence data).

Working with Data Owners on Processes to Obtain Data

Available data dictionaries were obtained and examined from each of the data sources selected in the environmental scan. Dataset contents were assessed to confirm that key variables required for sub-county estimate generation were present. Key variables included geographic designations (e.g., ZIP code, minor civil division) and demographic designations (e.g., gender, age group, insurance status, race/ethnicity) for aggregating data to sub-county level estimates, as well as measure parameters specified in definitions (e.g., *mother's age* for the teen pregnancy measure). This assessment allowed the research team to submit data requests to data owners that were explicit and comprehensive, to obtain the necessary data as efficiently as possible.

Data owners were essential to the project, not only by providing data, but also by answering technical questions, pointing out and clarifying data caveats, and particularly by supplying data suppression rules. Data owners often provided data accessing forms for requesters to complete, as well as formal data agreements to ensure the use of data was appropriate and aligned with the data request proposal. Data confidentiality rules were provided to data users so that estimates could be appropriately suppressed before being included in the final reports. As a condition for receiving certain data, the research team additionally needed to address data security, and ensure that access to raw data was limited only to appropriate staff who signed the respective data user agreements.

Measure Selection

Once data were obtained, the research team assessed the data and re-assessed the measure definitions to confirm which measures could be generated as proposed, which ones would require modification, and if any of the measures could not be generated. Among these were two measures for which the team proposed, and strongly justified modified definitions.

Two of the modifications were possible and appropriate due to NYSDOH having data for New York State that are not available for all states nationally. The first modification was to change *Teen births* to *Teen pregnancies*, with the justification being that pregnancies are more proximal to social determinants of health, safe sex behaviors, and other factors that public health practitioners aim to improve with local intervention. The second modification was to generate *Preventable hospital stays* for all patients rather than only for Medicare patients, with the justification being that generating data for all patients would provide a more representative indication of clinical and public health outcomes. The New York State hospital discharge database contained data for all age groups, and NYSDOH has used a national Prevention Quality Indicators #90 (PQI #90) among all adults (ages 18 years and older). Therefore, the County Health Rankings measure, "number of hospital stays for ambulatory-care sensitive conditions per 1,000 Medicare enrollees" (*Preventable hospital stays*), was substituted by the measure PQI #90.

NYSDOH did not have sub-county data for one County Health Rankings measure, *percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities*, exactly as it was defined. The research team assessed existing state-specific data sources and identified a measure that is related to housing affordability, "percentage of adults who report being always, usually, or sometimes stressed about having enough money for their rent or mortgage." Sub-county data for this measure were collected, and were available. Therefore, the available measure was selected instead.

Indicator Definition Adaptation

- Definitions have been adjusted for several measures to meet New York's data system needs as well as those of the statewide public health practitioner community:
    - Teen Birth Rate was replaced by Teen Pregnancy rate
    - Valid observations for rate of Low Birthweight calculation were defined as those with birthweights recorded between 100 and 8000 grams
    - Definition of Preventable Hospital Stays indicator was changed from '*preventable stays among Medicare enrollees / number of Medicare enrollees'* to '*all preventable stays / population*.'

The team decided to move forward with generating eleven measures:

1. Premature death
2. Poor mental health
3. Low birthweight
4. Adult smoking
5. Adult obesity
6. Food insecurity
7. Excessive drinking
8. Teen pregnancies
9. Preventable hospital stays
10. Injury deaths
11. Housing insecurity

Determining Units of Analysis

Defining Demographic and Geographic Units

The project required the team to select sub-county units for generating the measure estimates. Unit selection was based on the availability of the patient's geographic and demographic information in the data sources, in combination with the unit's utility for public health assessment. The selections were further guided by minimum sample sizes or numbers of events required for generating and sufficiently unsuppressed estimates.

For count data collected from vital statistics and hospitalization data, ZIP codes were the resulting sub-county geographic unit selected for all measures. For county sub-populations, race/ethnicity was selected as a demographic unit of analysis for all these measures, while other demographic unit selections varied by measure, and consisted of age group, Medicaid status, and education level.

For the survey data collected through an expanded, county level BRFSS, (eBRFSS), county subdivisions – referred to as minor civil divisions (MCDs) – were chosen as the geographical unit to use in calculating health measure estimates. MCDs, such as cities, towns, or reservations, are legally incorporated municipal corporations, providing services to their residents and empowered to tax property within their boundaries to raise revenue. There are 1,023 MCDs in NYS, including 932 towns, 62 cities, 14 Native American reservations, 10 undefined MCDs consisting entirely of water, and five town-village governments.

Population data from the American Community Survey (2009-2013 ACS Total Population) for both county subdivisions (and counties) were used to select 18 MCDs outside NYC for this pilot. MCDs were selected if they comprised more than 30 percent of the estimated county population. These MCDs included seven towns and 11 cities, and covered 17 counties. A list of the 18 MCDs and their associated counties is included in the table below.

#	Name of Minor Civil Division (County Subdivision)	County
1	Albany City	Albany
2	Amsterdam City	Montgomery
3	Auburn City	Cayuga
4	Binghamton City	Broome
5	Brookhaven Town	Suffolk
6	Buffalo City	Erie
7	Carmel Town	Putnam
8	Colonie Town	Albany
9	Cortland City	Cortland
10	Hempstead Town	Nassau
11	Owego Town	Tioga
12	Queensbury Town	Warren
13	Ramapo Town	Rockland
14	Rochester City	Monroe
15	Schenectady City	Schenectady
16	Syracuse City	Onondaga
17	Troy City	Rensselaer
18	Yonkers City	Westchester

To define these MCDs within the eBRFSS data file, the selected MCDs were assigned a group of respondent-level ZIP codes. The group of ZIP code assignments was made based on information within the 2010 Zip Code Tabulation Area (ZCTA) to County Subdivision Relationship file. Because ZCTAs were not coterminous with town and city boundaries, assignment of ZCTA to MCDs had to be based on percentage of population allocated. For this pilot, a ZCTA was assigned to an MCD if more than 50 percent of the population within the ZTCA resided within the MCD. The extent to which the population of a MCD was captured by the ZCTAs varied across the 18 MCDs used in the pilot and ranged from 68.5 percent (Schenectady City) to 100 percent (Amsterdam, Auburn, Buffalo, Troy and Cortland).

Summary Table of Sample Size Based on MCD to Zipcode Crosswalk

County Subunit Name	County	County Subunit ID#	County Subunit Estimated Population	% of County Estimated Pop	Est. eBRFSS Sample Size
Albany City	Albany	3600101000	98,142	32.1%	**290**
Amsterdam City	Montgomery	3605702066	18,425	36.8%	**223**
Auburn City	Cayuga	3601103078	27,571	34.6%	**250**
Binghamton City	Broome	3600706607	46,975	23.6%	**175**
Brookhaven Town	Suffolk	3610310000	486,868	32.5%	**149**
Buffalo City	Erie	3602911000	260,568	28.3%	**318**
Carmel Town	Putnam	3607912529	34,379	34.5%	**92**
Colonie Town	Albany	3600117343	81,908	26.8%	**195**
Cortland City	Cortland	3602318388	19,187	38.9%	**227**
Hempstead Town	Nassau	3605934000	761,975	56.7%	**275**
Owego Town	Tioga	3610755893	19,742	38.9%	**150**
Queensbury Town	Warren	3611360356	27,845	42.5%	**153**
Ramapo Town	Rockland	3608760510	128,336	40.7%	**183**
Rochester City	Monroe	3605563000	210,624	28.2%	**292**
Schenectady City	Schenectady	3609365508	65,990	42.6%	**171**
Syracuse City	Onondaga	3606773000	144,742	31.0%	**265**
Troy City	Rensselaer	3608375484	50,019	31.3%	**228**
Yonkers City	Westchester	3611984000	197,493	20.7%	**140**

Defining Time Periods for Health Indicator Estimates

Multiple years of data were combined to generate more stable estimates when the number of events for an indicator was small (such as rare conditions). Estimate stability is also affected by the selection of larger geographic or demographic units of analysis. Tradeoffs are therefore inherent in the unit selection process: granular analysis units and short time intervals reveal variation in health status between groups and facilitate targeted public health intervention, but also yield increasingly suppressed and unstable estimates as the sample sizes and numbers of events for each group decrease; broader analysis units and longer time intervals dilute or conceal variation in health status between groups and provide limited utility for targeted public health intervention, but they yield increasingly stable and unsuppressed estimates as the sample sizes and numbers of events for each group increases.

After a comprehensive assessment of measures, data sources, and geographic and demographic groups, the research team selected the following combinations of measures, data sources, years of data, and units of analysis:

Measure	Description	Data Source	Years	Level of Analysis
Premature death	Premature Death is the years of potential life lost before age 75 (YPLL-75). Every death occurring before the age of 75 contributes to the total number of years of potential life lost. For example, a person dying at age 25 contributes 50 years of life lost, whereas a person who dies at age 65 contributes 10 years of life lost to a county's YPLL. The YPLL measure is presented as a rate per 100,000 population and is age-adjusted to the 2000 US population.	New York State Vital Records	2009-2013	Race/ethnicity, ZIP code, county total
Poor mental health	Percentage of adults who reported that their mental health was poor or not good on at least 14 of the past 30 days. "Poor" and "not good" mental health days include days when there was stress, depression, and problems with emotions.	New York State 2013-2014 Expanded Behavioral Risk Factor Surveillance Survey (eBRFSS)	April 2013 - March 2014	Minor civil division (where data available), county total
Low birthweight	The percentage of births born weighing less than 2,500 grams (excludes births with unknown birthweight).	New York State Vital Records	2007-2013	Race/ethnicity, age group, Medicaid status, education, ZIP code, county total
Adult smoking	Percentage of adults who report smoking at least 100 cigarettes in their lifetime, and currently smoke on at least some days.	eBRFSS	April 2013 - March 2014	Minor civil division (where data available), county total
Adult obesity	Percentage of adults who are obese (i.e., body mass index greater than or equal to 30.0) based on self-reported weight and height.	eBRFSS	April 2013 - March 2014	Minor civil division (where data available), county total
Food insecurity	Percentage of adults who report being always, usually, or sometimes stressed about having enough money to buy nutritious meals.	eBRFSS	April 2013 - March 2014	Minor civil division (where data available), county total
Excessive drinking	Heavy or binge drinking is defined as: (a) consuming 5 (men) / 4 (women) or more drinks on an occasion during the past 30 days, or consuming greater than 2 (men) / 1 (women) alcoholic beverages per day in the past 30 days.	eBRFSS	April 2013 - March 2014	Minor civil division (where data available), county total

Teen pregnancies	Teen pregnancy rate per 1,000 female population aged 15-19 years. Pregnancies are the sum of the number of live births, induced terminations of pregnancies, and all fetal deaths.	New York State Vital Records	2011-2013	Race/ethnicity, ZIP code, county total
Preventable hospital stays	The number of preventable hospitalizations per 10,000 population aged 18+ years. This rate is age-adjusted to the 2000 US population. The Prevention Quality Indicators (PQIs) are a set of measures developed by the federal Agency for Healthcare Research and Quality for use in assessing the quality of outpatient care for "ambulatory care sensitive conditions." This rate is defined as the combination of the 12 PQIs that pertain to adults: (1) short-term complication of diabetes; (2) long-term complication of diabetes; (3) uncontrolled diabetes; (4) lower-extremity amputation among patients with diabetes; (5) hypertension; (6) congestive heart failure; (7) angina; (8) chronic obstructive pulmonary disease; (9) asthma; (10) dehydration; (11) bacterial pneumonia; (12) urinary tract infection. The PQIs estimate the number of potentially avoidable hospital admissions, and therefore a lower rate is desirable.	Statewide Planning and Research Cooperative System (SPARCS)	2011-2013	Race/ethnicity, age group, ZIP code, county Total
Injury deaths	Injury Deaths is the number of deaths from intentional and unintentional injuries per 100,000 population. Deaths included are those with an underlying cause of injury (ICD-10 codes *U01-*U03, V01-Y36, Y85-Y87, Y89).	New York State Vital Records	2009-2013	Race/ethnicity, age group, ZIP code, county total
Housing insecurity	Percentage of adults who report being always, usually, or sometimes stressed about having enough money for their rent or mortgage.	eBRFSS	April 2013 - March 2014	Minor civil division (where data available), county total

Prepare and Analyze Data

The research team prepared individual record-level data, first, by confirming that datasets received contained sufficient variables that were necessary for analysis due to their presence in measure definitions. The following steps are necessary for preparing and analyzing the data:

checking values for variables that are necessary for measure definitions (e.g., mother's age for the teen pregnancy measure; birthweight for the low birthweight measure) in the datasets; conducting univariate analyses; cleaning data; creating computed variables, if necessary; aggregating data by geographic and demographic levels; merging with population data; calculating measure estimates (percentages or rates); and then designing and populating a master dataset (Appendixes A-C: NYS Masterfile with accompanying data dictionary and codebook) structure that lends itself to the desired production workflow and data output.

Univariate analyses were subsequently conducted in order to ascertain frequencies and missing data for categorical variables; as well as distribution parameters (e.g., range, skew, kurtosis, and modality), missing data, and outliers for continuous variables.

Assessing distributions provided early indications of whether age-adjustment would skew the resulting estimates among small-population counties and county sub-populations. The distributions and the value ranges were also useful for identifying outliers, determining if they fell within plausible ranges (e.g., 'age=200 years' would be deemed implausible), and for determining whether identified problems were isolated occurrences or systematically present throughout the data. Data were then cleaned by converting variables to their appropriate data types (e.g., character to numeric), and suppressing invalid and implausible values.

Checking data distributions for each demographic variable also helped in determining appropriate groupings of values into sub-groups for displaying main measures by subpopulation groups. This was to ensure reasonable cell size for each subgroup to achieve statistical stability of the estimates in the majority of counties. For example, how demographic variable distributions informed the team's approach for combining ages and racial/ethnic groups for county subpopulations. Furthermore, for many measures, the rates are calculated using populations for certain geographic levels (county, ZIP code) and county subpopulations (age groups, race/ethnicity groups) as denominators. Therefore, assessing the availability of population denominator data is essential. This project utilized county population data from the US Census Bureau, and Nielson ZIP code population data. It is also important to evaluate the availability of population estimates by age group so that the age-adjusted rates can be calculated properly.

Studying and understanding the raw data, its variables, and what variables must be calculated or generated for the final product provides better information for the research team to consider, plan, and prepare in the early stages. In this project, the team worked with four very different data sources to generate eleven health measures. The final reports presented these measure estimates by state, county, and ZIP code levels, as well as by county subpopulation groups. It was important to anticipate and plan for what variables are needed to include in the master dataset that is used to populate the final reports when assessing different data sources and measures. The research team designed a mockup report with dummy data for all measures with possible visualizations such as graphs, tables, and maps to see what data/variables and information were needed for the master dataset to generate the report. From there the research team recognized that measure, unit of analysis (both demographic and geographic units

together), numerator, denominator, estimate value, confidence interval limits, stable (yes/no), and suppressed (yes/no) were the variables needed to sufficiently describe each output statistic. These variables were selected to structure the master dataset, with the aim of facilitating the desired production workflow and report output. The research team planned to automate production of the final PDF reports, one for each of New York's 62 counties, by using the master dataset as a single input for various SAS procedures.

After obtaining and validating the individual-record level data, the research team generated aggregated summary estimates for each measure, according to their respective definitions and units of analysis.

Assessing Data Validity, Outliers, Stability, and Suppression

After generating estimates (percentages or rates) at ZIP code level or by county subpopulation group, the research team conducted univariate analyses of the estimates. The univariate distributions were reviewed, and estimates that exceeded 90th percentile among values in the distribution for the respective measure were initially considered outliers, and therefore were manually examined and validated against county-level estimates.

The examination of outlying estimates was conducted carefully because the research team wanted to apply tertiary suppression, i.e., exclude erroneous data (e.g., coding errors, differentially-adjusted estimates) while avoiding exclusion of 'real' data with outlying values that reflect true health burden in the population. Extreme values were investigated individually. The research team had more confidence in estimates with large denominators. For age-adjusted estimates, the age distributions of the underlying populations were carefully checked as well.

In all cases, knowledge about community demographics provided valuable context to inform the team's determination to suppress or report extreme values. In one example in which community knowledge was applied was for a ZIP code level teen pregnancy rate, which was four times the statewide rate. This rate, without community knowledge, would be deemed implausibly high relative to nearby and statewide ZIP codes. However, the research teams determined the estimate to be valid because it had been independently corroborated by a community health assessment that had recently been conducted by a hospital and local health department.

After estimate distributions and outlying values were checked, the research team defined and applied criteria for flagging unstable estimates, and for suppressing estimates that could compromise individual confidentiality if reported. The suppression criteria were based on the rules that were set by specific data owner, which could be based on case/event counts or the sizes of estimates' underlying populations. The stability criteria were based on estimates' data types.

For count-related measures, estimates with relative standard error (RSE) greater than 30 percent should be considered unreliable/unstable (https://www.cdc.gov/nchs/data/bsc/bscpres_parker_january2015.pdf; https://health.ny.gov/diseases/chronic/ratesmall.htm). This usually occurs when there are fewer than 10 events in the numerator. The RSE is calculated by dividing the standard error of the

estimate by the estimate itself, then multiplying that result by 100. The RSE is expressed as a percentage of the estimate. Estimates with large RSEs are less reliable than those with small RSEs. All unsuppressed count measure estimates with RSE greater than 30 percent or with fewer than 10 events in the numerator were flagged as unstable by the research team in the final data reports.

For survey-related measures, estimates were considered unreliable/unstable when the width of the 95 percent confidence interval was greater than 20 percent (for percentage estimates) and/or the RSE is greater than 30 percent. All unsuppressed survey estimates meeting either of these criteria were flagged as unstable by the research team in the final data reports.

Results were suppressed (not shown on the final reports) when reporting could potentially have compromised individual confidentiality or increased the probability of identifying individuals who were reported with a health status related to a measure. Suppression rules vary depending on the data source. In New York State, the research team received the following suppression rules from data owners:

Data Source	Suppression Criteria
Survey data (eBRFSS)	Numerator <10 or denominator <50
Death data (Vital Records)	Denominator population <50
Birth data (Vital Records)	Denominator total births <30
Adolescent pregnancy data (Vital Records)	Denominator population <50
Hospitalization data (SPARCS)	Numerator cases <6

In addition to these primary suppression rules, the research team also applied secondary suppression to protect confidentiality in cases where estimates were suppressed for only one subgroup. By the 'pigeon hole principle,' a primary-suppressed estimate for a county subgroup could be calculated if each of the other subgroups is unsuppressed, and the county total is obtained. For example, in a county with three age groups, a1, a2, and a3, with a1 primary suppressed, and with A representing the county total for all three age groups combined, a1 could be calculated by subtracting the sum of the other subgroups from the county total:

$$a1 = A - (a2 + a3)$$

Therefore, in every instance when an estimate for only one subgroup was primary suppressed, the research team secondary suppressed the estimate for another subgroup.

Once primary and secondary suppression criteria were applied, and unstable estimates were flagged, the research team assessed the overall impact on suppression and instability on the aggregated data:

Percent of sub-county estimates that were suppressed, by measure*:

Premature death:	1.9%
Low birthweight:	8.5%
Teen pregnancies:	10.2%
Preventable hospital stays:	6.5%
Injury deaths:	0.4%

*Survey estimates did not meet suppression criteria

Percent of sub-county rates that were unstable, by measure:

Birth measures	
Low birthweight:	19.4%
Teen pregnancies:	20.8%
Death measures	
Premature deaths:	10.3%
Injury deaths:	47.8%
Hospitalization measures	
Preventable hospital stays:	2.3%
Survey measures**	
Adult obesity:	55.6%
Adult smoking:	33.3%
Excessive drinking:	33.3%
Food insecurity:	50.0%
Housing insecurity:	66.7%
Poor mental health:	55.6%

**Syracuse, Buffalo, and Rochester City were the only MCDs where stable estimates could be calculated for each of the 6 survey data measures

Unstable estimates by measure and minor civil division

Unstable MCD Estimates							
	Adult Obesity	Adult Smoking	Excessive Drinking	Food Insecurity	Housing Insecurity	Poor Mental Health	Grand Total
Albany City	0	0	0	1	0	0	1
Amsterdam City	1	0	0	1	1	0	3
Auburn City	1	1	0	1	0	1	4
Binghamton City	0	0	0	0	1	0	1
Brookhaven Town	1	1	1	1	1	0	5
Buffalo City	0	0	0	0	0	0	0
Carmel Town	1	1	1	1	1	1	6
Colonie Town	0	1	0	0	1	1	3
Cortland City	1	0	1	0	1	1	4
Hempstead Town	0	0	0	0	0	1	1
Owego Town	1	0	1	0	1	1	4
Queensbury Town	1	0	0	1	1	1	4
Ramapo Town	0	1	1	0	1	1	4
Rochester City	0	0	0	0	0	0	0
Schenectady City	1	0	1	1	1	0	4
Syracuse City	0	0	0	0	0	0	0
Troy City	1	1	0	1	1	1	5
Yonkers City	1	0	0	1	1	1	4
Grand Total	10	6	6	9	12	10	53

49% of estimates were unstable; However, the three largest upstate NY cities (Syracuse, Rochester and Buffalo) had stable estimates for all 6 measures.

Organizing and Aggregating Data for Generating Final Reports

With sub-county, county, and regional estimates generated and validated, the research team organized all of them into the master dataset that was structured to facilitate efficient production of county reports. Estimates in the master dataset were grouped and sorted by the following variables: measure, county, unit of analysis (e.g., *ZIP code*, *MCD*, *age group*, *race/ethnicity*), and unit value. The master dataset also had columns to record the years of data for each estimate, the underlying data type, whether or not numerators denominators and confidence intervals were to be reported, numerators (suppressed and unsuppressed), denominators (suppressed and unsuppressed), upper and lower confidence interval limits, and flags for suppressed and unstable estimates. This structure (see Table 3) enabled the research team to parameterize each of the desired reports, and batch process them from a single input dataset in SAS.

Master dataset structure

Variable	**Type**	**Description**
Indicator	Character	Name of measure
Unit_of_Analysis	Character	Level of data Demographic: (Age Group, Education, Medicaid Status, Race/Ethnicity) Geographic: (ZIP Code, Minor Civil Division, County, DSRIP Region, NYC/ROS, NYS)
Unit Value	Character	Specific sub-county population represented by that data with respect to the population category specified in *Unit_of_Analysis*
County	Character	If row corresponds with a county-, sub-county geographic-, or sub-county demographic- level estimate, then the *County* variable contains the respective county name. If row corresponds with a DSRIP region, New York City, Rest-of-State, or New York State, then the *County* contains the DSRIP region name, NYC, ROS, or NYS, respectively.
Suppressed	Character	Flag (s) for suppressed estimate numerators and rates.
Numerator	Numeric	Premature deaths: numerator = total premature deaths Low birthweight: numerator = total low birthweight births Teen pregnancies: numerator = total teen pregnancies Preventable hospital stays: numerator = total preventable stays Injury deaths: numerator = total injury deaths
Denominator	Numeric	Low birthweight: denominator = total live births Teen pregnancies: denominator = total females ages 15-19
Rate	Numeric	Premature deaths: rate = age-adjusted YPLL per 100,000 Poor mental health: rate = percent of adults that report poor mental health Low birthweight: rate = percentage of live births with low birthweight Adult smoking: rate = percent of adults that report smoking Adult obesity: rate = percent of adults that report BMI > 30 Food insecurity: rate = percent of adults that report food insecurity Excessive drinking: rate = percent of adults that report drinking excessively Teen pregnancies: rate = teen pregnancies per 1,000 females ages 15-19 Preventable hospital stays: age-adjusted preventable hospitalizations per 10,000 adults (Note: age group rates are crude, not age-adjusted) Injury deaths: rate = injury deaths per 100,000 Housing insecurity: rate = percent of adults that report housing insecurity
Unstable	Character	Flag (*) for unstable rates. (see report Methods section for stability criteria)
Lower 95%CI Limit	Numeric	95% confidence interval for survey measures based on eBRFSS data
Upper 95%CI Limit	Numeric	

Design Format for Presenting Analyzed Results

The selected format for the reports was PDF. This was chosen over Microsoft Word, Rich Text Formats, and others because PDFs are relatively small in file size and conducive for digital distribution, they are readable on multiple operating systems, and they are difficult for end users to accidentally modify. The research team decided to produce 62 PDF reports (one per NYS county) rather than eleven reports by measure, or two reports by data type (i.e., one for count data, and one for survey data). This format was selected because the project's focus on public health practice; it was determined that a specific report for each county would allow local practitioners to have all of their data in one place, as opposed to the alternatives, which would have required them to use multiple reports in order to access their data.

The PDF reports each contained vertical bar graphs, data tables, and choropleth maps for the 11 selected sub-county measures (see all reports: http://www.nysacho.org/i4a/pages/index.cfm?pageID=3810). All of the PDF reports, including their respective data visualizations, were generated and compiled using the SAS software (Appendixes D-F: 'County rankings.sas', 'chrmacros.sas', 'pa_formats_2.sas7bdat'). Different SAS procedures were used such as PROC REPORT, PROC GCHART, and PROC GMAP to display tables, charts and maps in the final county report. The SAS MACRO facility was used to develop a program to perform the do-loop to automate the process for generating the 62 individual county reports consistently. The research team improved the scalability and efficiency of the process by implementing SAS macros together with a reference file (Appendix G: 'County_Ranking_reference_5_04_2016.csv'). The reference file includes all of the information and specifications that the SAS program uses for formatting the PDF output.

For the county maps by ZIP code, the research team applied a three-color scale based on quartile ranges of estimates for each measure. ZIP codes with estimates in the first and second quartiles were shaded yellow, ZIP code estimates in the third quartile were shaded light green, and ZIP code estimates in the fourth quartile were shaded dark blue. This scale was applied to visually highlight high-burden areas. The research team selected the color palette from Colorbrewer (http://colorbrewer2.org/) because the colors effectively contrast against one another (even in gray scale), and because the palette is colorblind safe, print friendly, and photocopy safe.

All aesthetic decisions were made with the aim of improving the utility of the reports for end-users. This included: clarity of wording for titles, labels, and footnotes; hyperlinks from the table of contents to bookmarks throughout the reports to facilitate easy navigation; as well as font choices to improve legibility. Additional factors were carefully considered to improve the utility of the choropleth maps, including the legend design, the sizes and placement of text labels, and the use of insets for 'busy' maps (e.g., maps of counties ZCTAs that vary greatly in size).

Soliciting End-users' Input, and Incorporating Their Feedback into the Project

To assess the usefulness, gaps, and opportunities for improvement of the sub-county data reports prior to their public release, the research team conducted a focus group with staff from local health departments (LHDs), hospitals and regional public health planning organizations who would potentially use the report. Specifically, the proposed focus group collected qualitative end-user data about: (1) participants' overall impressions of a mock-up report, its presentation, and its utility for local public health planning; (2) whether the report lends itself to accurate interpretation; and, (3) opportunities for improving the report (e.g., via formatting changes, description revisions). An IRB application was approved by NYSDOH IRB as exempt from full review.

To prepare for the focus group, invitations were sent via email to liaisons from LHDs, hospitals and regional public health planning organizations to determine interest in having a staff person from their organization participate in the focus group. If the liaison was interested in having their organization be represented, they were asked that they forward a link, contained in the email, to an electronic interest survey to the appropriate staff representative. It was specified that staff representatives should have roles that involve gathering or analyzing health-related data, participating in conducting community health needs assessment and planning, or participating in monitoring and evaluating interventions. Staff representatives then completed the survey, which described to respondents the study background, informed consent, and acceptance or declination of focus group participation.

Of the staff that indicated interest in attending the focus group, a quota sampling methodology was used to ensure that the final sample of focus group participants included 3 large LHDs, 3 small LHDs, 3 hospitals, and 3 regional public health planning organizations from a sampling frame that includes: all LHDs in New York State except for the New York City Department of Health and Mental Hygiene, all non-profit hospitals in New York State, and any regional public health planning organizations that have participated in the New York State Prevention Agenda such as Population Health Improvement Program contractors (PHIPs). 'Large LHDs' were defined as any LHD that is located in the 17 counties for which this project has generated minor civil division-level data; all other LHDs we define here as 'small LHDs.'

The focus group was conducted virtually using WebEx. The focus group was approximately one hour in duration, with 10 participants (two selected participants dropped out). The focus group facilitator read consent information, and explained that the session would be recorded for later transcription and analysis.

Information that was collected from focus group participants was non-personal, and related solely to their impression of the mock-up reports as subject matter experts for their organizations. To ensure anonymity of the focus group, all participants were assigned a number. The meeting organizers used WebEx to ask closed ended poll questions (including demographic questions) that were not visible to other participants. The recorded webinar was transcribed without use of personal identifiers, so that future reports will use de-identified aggregate-level data. Access to the recordings and transcriptions remains limited to the research team.

Transcripts of the focus group were reviewed and coded by the research staff. The end-user feedback provided valuable insights, which the research team subsequently incorporated

into the final reports. When asked whether a report containing measures with mostly suppressed estimates was still useful, the end-users indicated that they preferred to receive available data, even when much of it was suppressed. They also expressed their preference for data to be made available in multiple formats, in addition to the PDF reports.

Evaluation of Project Successes and Challenges

A pilot evaluation is ongoing, but it began concurrently with the project initiation by tracking the completion of major milestones from the originally proposed work plan (e.g., production of final reports, promotion to a public website, and presentation of a technical assistance webinar for end-users).

Ongoing evaluation activities now focuses on measuring outcomes. Data sources for this include page-views of the webpage where the reports are posted, attendance totals from scheduled webinars, as well as qualitative findings from key informant interviews with LHD staff and other end-users. The research team will also track the number of periodically updated community health assessments and community health improvement plans/community service plans published by hospitals and local health departments that cite or include the sub-county data produced by the New York pilot project.

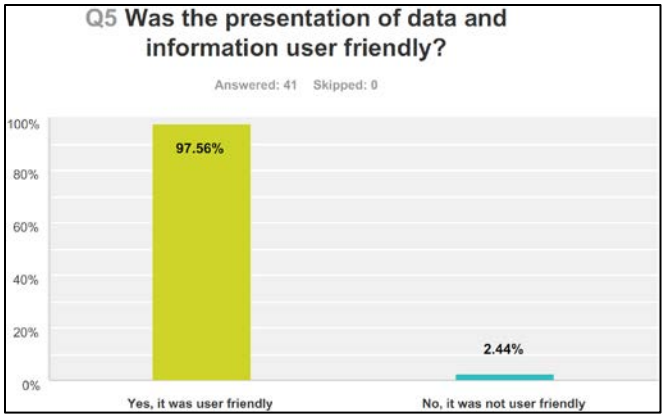
Results

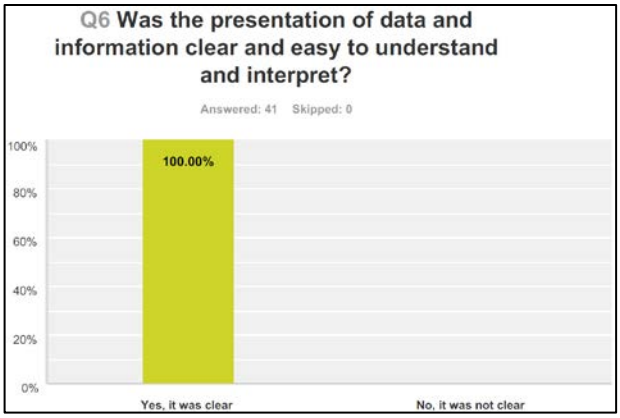
Products

The research team produced 62 PDF reports (one per New York State county), each containing state, regional, county, and sub-county level data for 11 measures.

Collection and Implementation of End-user Feedback

Shortly after the final reports were publicly released, end-user feedback was ascertained via an online survey. Forty-one participants responded, of which 98% indicated that the reports' presentation of data and information was user-friendly, and 100% indicated that the presentation of data and information was clear and easy to interpret:





Forty of the respondents also indicated that they were using the PDF reports for at least one purpose:

Product Distribution, Communication/Marketing, and Provision of Technical Assistance

All 62 PDF reports are available on the New York State Association-County Health Officials website since May 2016 (http://www.nysacho.org/i4a/pages/index.cfm?pageID=3810). The reports were also emailed directly to their respective county's local health department prior being publicly released. A public webinar with participation of staff from local health departments, hospitals, regional health planning groups, and other partners was conducted in June 2016. During the webinar, the research team introduced the reports, and provided guidance to support the participants in using the measures and included data for a variety of purposes – including community health needs assessments, and identifying high-burden areas and populations.

Strengths

- Access to a variety of health datasets, facilitated through existing relationships with data owners.
- Implementation of report templates, SAS, and an automated production process.
- Application of primary, secondary, and tertiary suppression criteria.

Technical Lessons Learned

- Challenges in selecting, analyzing and presenting measures:
  - It is challenging to find balance between selecting important health measures and ensuring the outcome of interest bears reasonable sample size at the sub-county level for the majority of the sub-county areas.
- Classifying outliers as "real estimates," or as products of "data issues" (i.e., data entry errors, skewed adjustment).
- Determining the optimal balance among competing factors for producing sub-county estimates for local public health planning: lower suppression/higher stability, fewer years of combined data, smaller geographic/demographic units of analysis.
- Involve key stakeholders or audiences in early stages of planning and implementation to improve final products and overall project efficiency. Feedback from end users' can validate the utility of intended products during their development so that rework can be avoided. However, it is important to balance stakeholders' needs and desires with technical feasibility. For example, stakeholders may request census block- or street level- data to facilitate their community health planning; however, feasibility may require less granular reporting in order to derive unsuppressed estimates.

Sustainability and Next Steps

This project provided an exercise for the New York State research team to specifically examine sub-county level data analysis and presentation with the aim of effectively supporting public health activities at the local level. It also allowed the research team to interact with key audience groups that use these data to gain feedback on how to present results and visualize main estimates in the most effective way for them.

Even though, this is not a venue to provide New York State sub-county data on an ongoing basis, the experience and lessons learn from this project were valuable to the research team to improve the quality and presentation of sub-county level data presented in a sustainable NYSDOH dashboard application: https://health.ny.gov/preventionagendadashboard

Acknowledgements

New York State Department of Health

Trang Nguyen, MD, DrPH, MPH
Deputy Director, Office of Public Health Practice

Ian Brissette, PhD
Director, Bureau of Chronic Disease Evaluation and Research

Isaac Michaels, MPH
Evaluation Specialist, Office of Public Health Practice

Yunshu Li, MS
Public Health Informatics Specialist, Office of Public Health Practice

Mycroft Sowizral, PhD
Research Scientist, Bureau of Chronic Disease Evaluation and Research

Anne McCarthy, MPH
Research Scientist, Bureau of Chronic Disease Evaluation and Research

Asante Shipp-Hilts, DrPH
Project Coordinator, Office of Public Health Practice