2010 Community Needs Assessment: Secondary Data Analysis Methodology

Secondary Analysis Conducted by:

Mathew Schmidtlein, PhD
Assistant Professor
California State University
schmidtlein@csus.edu

2010 Community Needs Assessment Project Lead:

Valley Vision, Inc.
2320 Broadway
Sacramento, CA 95818
www.valleyvision.org

August, 2010

The secondary data analysis for the 2010 Community Needs Assessment consisted of three parts: data collection, data processing, and data analysis. The data collection component consisted of identifying and obtaining relevant demographic and public health data for the study area. Data processing involved standardization and other steps needed to make the data useful. Data analysis involved the development and implementation of a strategy that allowed meaningful patterns to be identified within the large set of data collected. The detail of each of these steps is given below.

Data Collection

Data collection focused primarily on obtaining relevant demographic and public health data for the 106 ZIP codes within the study area. Data were collected at the ZIP code level for the years 2006, 2007, and 2008. Three main data sources were used: Emergency Department (ED) Visit and Hospitalization data were obtained from the California Office of Statewide Health Planning and Development (OSHPD); birth and mortality data were obtained from the Birth and Death profiles published by the California Department of Public Health (CDPH); and demographic data were obtained from Thomson Reuters. The OSHPD data gives the number of ED Visits or Hospitalizations that had specific ICD9/E codes associated with a primary or other condition. The tables below identify the specific variables obtained from each source:

OSHPD ED Visit and Hospitalization Data

Variable Name	Description	Years Available	ICD9/ E Codes
H_AST	Asthma related hospitalization	2006-2008	ICD9: 493-494
H_CAN	Cancer related hospitalization	2006-2008	ICD9: 140-149, 150-159, 160-165, 170-176, 179-189, 190-199, 200-209
H_DIA	Diabetes (type II) related hospitalization	2006-2008	ICD9: 250
H_HIP	Hip Fracture related hospitalization	2006-2008	ICD9: 820
H_HTD	Heart Disease related hospitalization	2006-2008	ICD9: 410-417, 428, 440, 443, 444, 445, 452
H_HYP	Hypertension related hospitalization	2006-2008	ICD9: 401-405
H_INJ	Unintentional Injury related hospitalization	2006-2008	ICD9: 800-950; E1-E4
H_MNH	Mental Health Disorder related hospitalization	2006-2008	ICD9: 290-299; 300-316
H_STK	Stroke related hospitalization	2006-2008	ICD9: 430-436, 438
ED_AST	Asthma related emergency department visit	2006-2008	ICD9: 493-494
ED_CAN	Cancer related emergency department visit	2006-2008	ICD9: 140-149, 150-159, 160-165, 170-176, 179-189, 190-199, 200-209
ED_DIA	Diabetes (type II) related emergency department visit	2006-2008	ICD9: 250
ED_HIP	Hip Fracture related emergency	2006-2008	ICD9: 820

	department visit		
ED_HOM	Homicide related emergency department visit	2006-2008	E960-E969
ED_HTD	Heart Disease related emergency department visit	2006-2008	ICD9: 410-417, 428, 440, 443, 444, 445, 452
ED_HYP	Hypertension related emergency department visit	2006-2008	ICD9: 401-405
ED_INJ	Unintentional Injury related emergency department visit	2006-2008	ICD9: 800-950; E1-E4
ED_MNH	Mental Health Disorder related emergency department visit	2006-2008	ICD9: 290-299; 300-316
ED_STK	Stroke related emergency department visit	2006-2008	ICD9: 430-436, 438
ED_SUI	Suicide related emergency department visit	2006-2008	E950-E959

CDPH Mortality and Birth Data

Variable Name	Description	Years Available	ICD9/ E Codes
HTD	Mortality from Diseases of Heart	2006-2008	ICD10: I00-I09, I11, I13, I20-I51
CAN	Mortality from Malignant Neoplasms (Cancer)	2006-2008	ICD10: C00-C97
STK	Mortality from Cerebrovascular Disease (Stroke)	2006-2008	ICD10: I60-I69
CLD	Mortality from Chronic Lower Respiratory Disease	2006-2008	ICD10: J40-J47
INJ	Mortality from Unintentional Injuries (Accidents)	2006-2008	ICD10: V01-X59, Y85-Y86
ALZ	Mortality from Alzheimer's Disease	2006-2008	ICD10: G30
DIA	Mortality from Diabetes Mellitus	2006-2008	ICD10: E10-E14
PNF	Mortality from Influenza and Pneumonia	2006-2008	ICD10: J09-J18
LIV	Mortality from Chronic Liver Disease and Cirrhosis	2006-2008	ICD10: K70, K73, K74
SUI	Mortality from Intentional Self Harm (Suicide)	2006-2008	ICD10: U03, X60-X84, Y87.0
HYP	Mortality from Essential Hypertension & Hypertensive Renal Disease	2006-2008	ICD10: I10, I12, I15
HOM	Mortality from Assault (Homicide)	2006-2007	ICD10: U01-U02, X85-Y09, Y87.1
OTH	Mortality from All Other Causes	2006-2008	ICD10: Residual Codes
Total Deaths	Total Deaths	2006-2008	
M	Male Deaths	2006-2008	
F	Female Deaths	2006-2008	
<1	Deaths of individuals under 1 year old	2006-2008	
1-4	Deaths of individuals from 1 to 4 years old	2006-2008	
5-14	Deaths of individuals from 5 to 14 years old	2006-2008	
15-24	Deaths of individuals from 15 to 24 years old	2006-2008	
25-34	Deaths of individuals from 25 to 34 years old	2006-2008	
35-44	Deaths of individuals from 35 to 44 years old	2006-2008	
45-54	Deaths of individuals from 44 to 54 years old	2006-2008	
55-64	Deaths of individuals from 55 to 64 years old	2006-2008	
65-74	Deaths of individuals from 65 to 74 years old	2006-2008	

Variable Name	Description	Years Available	ICD9/ E Codes
75-84	Deaths of individuals from 75 to 84 years old	2006-2008	
85+	Deaths of individuals 85 years old and over	2006-2008	
Unk	Deaths of individuals with unknown age	2006-2008	
Births	Total births	2006-2008	
<1500	Births with weight under 1500 grams	2006-2008	

Thomson Reuters Demographic Data

Variable Name	Description	Years Available
Poverty 65+	Percentage of households below poverty line, with the head of household age 65 or more	2005,2007-2008
Poverty Children	Percentage of families with children under 18 below poverty line	2005,2007-2008
Poverty Single w/kids	Percentage of single female-headed families with children under 18 below poverty line	2005,2007-2008
No High School Diploma	Percentage of population over 25 without a high school diploma	2005,2007-2008
Minority	Percentage of population that is minority (including Hispanic ethnicity)	2005,2007-2008
Limited English	Percentage of population over age 5 that speaks English poorly or not at all	2005,2007-2008
Unemployed	Percentage of population in the labor force, ages 16 or more, without employment	2005,2007-2008
Uninsured	Percentage of population without health insurance	2005,2007-2008
Renting	Percentage of households renting their home	2005,2007-2008
Population	Percentage of households below poverty line, with the head of household age 65 or more	2005,2007-2008

Additional age stratified population estimate data at the ZIP code level for 2004-2008 were obtained from GeoLytics. Age stratified population data for 2000 at the ZIP code, county, and state level were obtained from the 2000 Census. Age stratified and total population data at the county and state level for 2008 were obtained from the American Community Survey 2008 1 year estimate.
A number of additional variables were collected at the county level. While these variables were not available at the ZIP code level, they were nevertheless included because of their importance in understanding regional health patterns. These variables are listed in the following table:

Point-in-time homeless population estimates	Admission to programs for drug and alcohol treatment	Percent days with air quality unhealthy for sensitive Individuals
HIV/AIDS Incidence by Race/Ethnicity	Age Adjusted Total Cancer Incidence by Race/Ethnicity	Age Adjusted Cancer Incidence by Sex and Type

Data Processing

A number of steps were undertaken to convert the public health and demographic data into forms suitable for a comparative analysis across the entire study area. These steps included the conversion of the raw count data in to standardized rates; the calculation of health indicators and a demographically based health index; and the calculation of variables at the county and state level to allow the comparison of ZIP code level rates to benchmarks. The specific details of each step are given below.

Rate Calculation

Rates are preferable to raw counts of data in comparative analyses because they reduce the impact of changing population sizes between study units. The Thomson Reuters demographic data was obtained in rate form. All mortality, ED visit, and hospitalization variables for all three years (2006-2008) were standardized based on total ZIP populations, so in the end the rates represented cases per 10,000 people[1].

Because the total population was so low for some ZIP codes in the study area, the reliability of the calculated rates was influenced by the small number problem. This means that areas with small populations may appear to have significantly higher rates than would be expected simply because the denominator is so small. To correct for this problem, the calculated rates were smoothed using empirical bayes (EB) smoothing whenever possible[2]. In this approach, rates for ZIP codes with small populations are adjusted to make them more similar to the expected average of the study area. Rates for areas with higher populations are adjusted less. This decreases the likelihood that a spuriously high rate will result from the small number problem.

Calculation of Health Indicators

In addition to the rates calculated for the ED, hospitalization, and mortality variables, a number of more complex variables though to represent different aspects of overall health were calculated. These include low birth weight, infant mortality rate, age adjusted mortality, and life expectancy at birth.

All health indicator variables were calculated using data from 2004-2008, and are based on EB smoothing. The combination of using data over a larger period of time, and using the smoothing approach, further reduced the impact of the small number problem in the calculation of the health indicators.

[1] When no data was reported for the raw count data, no rate value was calculated.
[2] When it was not possible to smooth the rate using EB smoothing, the un-smoothed rate was used.

Low birth weight is calculated as the total number of births with baby weights below 1500 gram divided by the total number of births[3], and is reported as the number of cases per 10,000 births. Infant mortality rate is calculated as the number of deaths in the population under 1 year old divided by the total number of births. It was also calculated using data from 2004-2008, and reported as cases per 10,000 births.

Age adjusted mortality involved dividing the age stratified mortality data obtained from CDPH by age stratified population and standardizing those values to the U.S. 2000 Standard Population. This is accomplished by multiplying the mortality rate (per 10,000) for each age group by the appropriate number given in Klein and Schoenborn[4], and adding the results of all age groups. The result is a number that represents the number of deaths per 10,000 people adjusted to match a standard age profile. One of the challenges to this approach was that the age stratified data available at the ZIP code level from 2004 to 2008 did not have the same interval widths as the age stratified mortality data. Specifically, the age stratified mortality listed deaths for the ages under 1 and from 1 to 4, while the age stratified population data only listed population with age less than 5. To overcome this issue, the population under 1 year old was estimated using two main strategies.

First, the ZIP code level population under 1 year old was estimated for ZIP codes from 2004-2008 using parameters generated from a regression model. Data on the population under 1 year old was obtained for each ZIP code from the 2000 census. A regression model was fit comparing total births for each ZIP code (from CDPH) in 2000 to the population under 1 in 2000. The total number of births in each year from 2004 to 2008 were then entered into the model to estimate the population under 1 year old in each year. This number was then subtracted from the variable reporting the number of people under 5 years old to estimate the population from 1 to 4.

In some cases the population under age 1 estimated according to the method outlined above was equal to or greater than the total population under 5 years old. In these instances, the population under the age of 1 was estimated using an alternative method. The average proportion of the population under the age of 5 that was also under the age of 1 was calculated for all ZIP codes in 2000. If the preceding approach resulted in an estimate of the population under 1 equal to or greater than the population under 5, this proportion was multiplied by the ZIP codes' population under 5 to yield the estimate of the population under 1. This second approach was also used to estimate the population under 1 for counties and for the state in 2008.

The final health indicator variable calculated was life expectancy at birth. This was calculated by first calculating the EB smoothed age stratified mortality rate (the percentage of population in each age strata that died). These rates were then used to calculate a period life table using a modification of the

[3] CDPH data masked birth information from ZIP codes with fewer than 5 births. For the health indicator analysis, it was assumed that there were 0 births for a ZIP code that had no data reported. This same approach was applied in the calculation of low birth weight, age adjusted mortality rate, and life expectancy at birth.

[4] Klein RJ, Schoenborn CA. Age adjustment using the 2000 projected U.S. population. Healthy People Statistical Notes, no. 20. Hyattsville, Maryland: National Center for Health Statistics. January 2001.

demogR[5] library of the R[6] statistical software package. The demogR package was modified to allow mortality rates to be entered instead of deaths and population figures so that the life table could be built using EB smoothed death rates, again to reduce the impact of the small number problem. Life expectancy at birth was extracted from the resulting life table.

Community Health Vulnerability Index

The next step in the data processing phase of the analysis was the creation of the Community Health Vulnerability Index (CHVI). This index was based largely on the Community Need Index (CNI) created by Catholic Healthcare West and Thomson Reuters[7]. The CNI (and the CHVI) is based on nine socio-economic/demographic variables. These variables correspond with conditions that are broadly recognized in the public health literature as contributing to poor health outcomes. The variables are shown in the table below:

Percentage of households below poverty line, with the head of household age 65 or more	Percentage of population over age 5 that speaks English poorly or not at all
Percentage of families with children under 18 below poverty line	Percentage of population in the labor force, ages 16 or more, without employment
Percentage of single female-headed families with children under 18 below poverty line	Percentage of population without health insurance
Percentage of population over 25 without a high school diploma	Percentage of households renting their home
Percentage of population that is minority (including Hispanic ethnicity)	

To create the CHVI, each of these variables was re-scaled so that the maximum value in the study area equaled 1 and the minimum equaled 0. The variables were then summed with equal weights, and the resulting values were converted into ranks, where low values represent areas with higher health vulnerability, and higher values represent areas with lower health vulnerability (so that a rank of 1 equals the most vulnerable zip code).

County and State Level Variables

Finally, rates for the ED, hospitalization, and mortality data, as well as low birth weight, infant mortality rate, age adjusted mortality, and life expectancy, were calculated for El Dorado, Place, Sacramento, and Yolo counties, as well as for the entire State of California. This was done so that they could serve as benchmarks against which to compare the ZIP code level rates and indicators. In order to allow for a more direct comparison, the county and state rates were based on an aggregation of ZIP code level data. Because of this, the "county" rates will not correspond exactly with county rates that may be calculated elsewhere. This is because the county rates here are calculated based on the rates for the ZIP codes associated with the counties, and ZIP code boundaries cross county lines. This means that the

[5] Jones, J. (2007). demogR: Analysis of age-structured demographic models. R package version 0.4.2.
[6] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
[7] Roth, R. and Barsi, E. 2005. The "Community Need Index." *Health Progress*. 86(4):32-38

county rates calculated here are approximations. The ZIP codes associated with each county are listed below:

County	Associated ZIP Codes
El Dorado	95619 95623 95633 95634 95635 95636 95651 95664 95667 95672 95682 95684 95709 95720 95721 95726 95762 96142 96150 00051 95614
Placer	95631 95648 95650 95658 95661 95663 95677 95678 95681 95701 95703 95713 95714 95715 95717 95722 95724 95746 95747 95765 96140 96141 96143 96145 96146 96148 95602 95603
Sacramento	95615 95621 95624 95626 95628 95630 95632 95638 95641 95652 95655 95660 95662 95670 95673 95683 95690 95693 95742 95757 95758 95814 95815 95816 95817 95818 95819 95820 95821 95822 95823 95824 95825 95826 95827 95828 95829 95830 95831 95832 95833 95834 95835 95837 95838 95841 95842 95843 95864 95608 95610 95836
Yolo	95616 95618 95627 95637 95645 95653 95679 95691 95694 95695 95698 95776 95937 95605 95606 95607 95612

The data sources and methodology used to calculate the county and state variables match that used to calculate the ZIP code level variables with a few exceptions. The small number problem was not an issue, so the steps used to address it in the ZIP code level analysis were not needed. So EB smoothing was not used in the calculation of county and state level rates, and the health indicator variables were calculated using only 1 year's worth of data. Additionally, as noted earlier, the population under 1 year old was estimated based on the proportion of the population under 5 years old that was also under 1 year old in 2000.

Data Analysis

The final stage of the secondary data analysis is the actual analysis of the processed data. The analysis began by identifying the top and bottom 10 percent of ZIP codes based on the CHVI variable. These ZIP codes were treated as two groups: the lowest health vulnerability group and the highest health vulnerability group. A series of boxplots were constructed that compared the values of the two groups on each ED, hospitalization, mortality, and health indicator variable. Finally, independent samples *t*-tests were used to determine if a statistically significant difference existed between the high and low CHVI groups.

For More Information

For more information on the 2010 Community Needs Assessment, visit www.healthylivingmap.com, or contact:

Valley Vision
2320 Broadway
Sacramento, CA 95818
916-325-1630
www.valleyvision.org