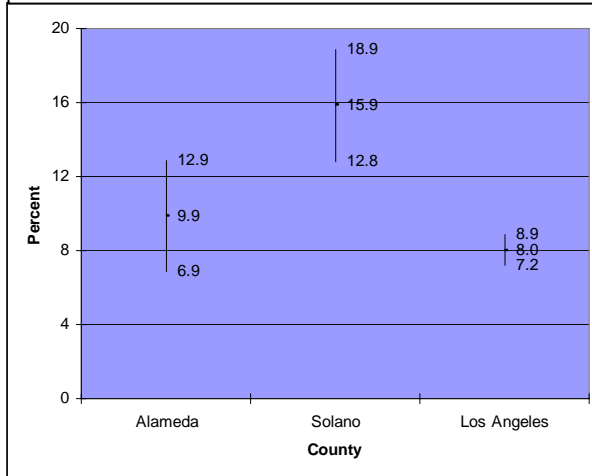




Data User Guideline #1:
 The California Health Interview Survey
 August 5, 2003
 Eric M. Roberts, MD PhD

Asthma Prevalence among Children, with confidence intervals (CHIS 2001 data)



Questions that we hear...

- Based on these findings, does Solano County really have the highest asthma prevalence in the Bay Area?
- Since the figures for LA County are so low, should we no longer be concerned about asthma there?

One of the most widely publicized events in the field of asthma advocacy in recent memory has been the release of the California Health Interview Survey (CHIS) data. Many around the state are using these data for planning and resource allocation, so we must be careful to use this data source appropriately.

In this User Guideline, we will discuss what CHIS is, and where it came from. We will review the strengths of the survey, its weaknesses, and (using a concrete example) what we can do to make sure we use the data appropriately.

We will also note some logistical issues in actually using the CHIS dataset.

What is CHIS?

The California Health Interview Survey (CHIS) was conducted in 2001 by the University of California Los Angeles Center for Health Policy Research, in cooperation with the California Department of Health Services and the Public Health Institute. CHIS was a telephone survey intended to answer questions about the health and health care of Californians. A total of 73,821 people were surveyed, including 5,801 adolescents and 12,592 children below age 12 (in which case adults—usually parents—were interviewed about the children in question).

The Good News...

There are several things that make CHIS a truly exciting and useful project. The survey is based on the highly comprehensive instrument used in the National Health Interview Survey (NHIS), making many of the items comparable to nationwide statistics that are very widely cited. Beyond this, the survey instrument was augmented with input from a number of stakeholders to make it useful for specific research projects and policy questions.

The CHIS staff also took a further step to increase the survey's usefulness. Nationwide surveys such as NHIS generally only provide information about the most numerous racial and ethnic groups in the country, and frequently data are only available for three groups (non-Hispanic whites, non-Hispanic blacks, and Hispanics). Given California's high degree

of diversity, CHIS conducted special sampling to be able to report data on Native Americans, Asians, and Pacific Islanders as well. The survey was conducted in several languages (English, Spanish, Mandarin, Cantonese, Vietnamese, Korean, and Khmer). Sampling was conducted with the goal of producing statistics for every California county or (for thinly populated regions) geographic area.

...and the Limitations

As impressive as the CHIS enterprise is, no dataset comes without limitations, and there are two particularly important ones to keep in mind.

The first of these is the response rate. In spite of the best efforts, only 33% of the people asked actually completed the survey for children. In combination with the fact that interviews were conducted exclusively by telephone, this fact raises the possibility of **bias** in the sample, which means that the group of people answering the survey could be different (for example, with respect to income level or cultural background) from the group that refused to participate. The higher the response rate, the less one needs to worry about bias actually changing your results. With a response rate this low, researchers begin to have anxiety about the accuracy of findings, particularly if they contradict what other people have found in other places using other methods.

The other caveat has to do with sample size. Although one of the goals was to provide information at the county level, the samples in individual counties were sometimes rather small, with the result that estimates have relatively large confidence intervals.

For a complete explanation of confidence intervals, please see the Supplemental Data User Guideline on the topic. In brief, while a survey may provide an **estimate** of, say, the prevalence of asthma symptoms, it is often more important to note the **range** of values within which we are reasonably confident that the **real** prevalence falls. This range is the confidence interval.

The issue of confidence intervals is not unique to CHIS; confidence intervals always get larger when the sample is small no matter what the project. Large confidence intervals can be frustrating, however. For example, we can't actually say for sure whether the childhood asthma prevalence in Solano is really higher than in Alameda because of confidence intervals that are too wide.

What to do

There are two actions, therefore, that we should all take when using survey data, and these are particularly relevant to the case of CHIS. The first is to always report findings in the context of previous information from other sources. Often there is no way to get rid of bias, although it can be noted and sometimes quantified. Presenting new findings alongside other information that may be contradictory gives the reader a chance to consider how the new findings should be interpreted.

The second action is equally important: always report rates **along with confidence intervals**. In a sense, the confidence interval can be more important than the estimated rate itself. These practices are demonstrated with the example below.

An Example: Childhood Asthma Prevalence

For example, the estimated childhood asthma prevalence for each

California County is presented along with 90%-confidence intervals on the CHIS website. Data for three counties are shown in the above graph.

How dependable is the information? We have to consider the possibility of sample bias, although we have no way to correct it. We can, however, think about what we already know about asthma in California and the United States. For example, CHIS finds a childhood asthma prevalence in Los Angeles County of 8.0% (90%-CI: 7.2,8.9), which is clearly lower than the state average. The rate of childhood asthma hospitalizations in the County for 1995-97, however, has been 242 per 100,000 (95%-CI: 239-246). This is 12% **higher** than the state average, a difference that also happens to be statistically significant.

Does this mean that CHIS data are wrong? Not necessarily—among other explanations, these findings may speak to an under-diagnosis of asthma among LA County children. What would be wrong, however, would be to use the CHIS data to say that LA County has less of an asthma problem than the rest of the state.

Another temptation is to compare all of the counties based on their estimated rates, as with Alameda (9.9%) vs. Solano (15.9%), to use the earlier example. The comparison is striking, since the relative difference between the two is about 60%. It turns out, however, that the confidence intervals overlap [(6.9,12.9) and (12.8,18.9), respectively]. Is there a difference? If we ran the same survey again would Solano still be higher than Alameda? It might, but we cannot tell from this information.

All together, 7 counties had childhood asthma rates with confidence intervals that did not overlap with the state average (4 were higher than average and

3 were lower). Since confidence intervals are not the perfect way to judge statistical significance, some other counties may be significantly higher or lower than average as well, but we can only speak with certainty about these 7.

How to Access CHIS

A public-use version of the CHIS dataset is available for download through the CHIS website (www.ucla.edu), with which one can conduct state-level analyses only. In order to preserve confidentiality, analyses making use of geographic identifiers require an application to the UCLA Data Access Center. The staff there are then able to facilitate on-site access to the complete dataset for more detailed work.

Because of the complex sample design employed in CHIS, special statistical methods are required in order to generate valid statistics (including confidence intervals). For this reason, it is inadvisable to use standard modeling techniques (t-tests, multiple regression) when using these data. Specialized procedures to account for the CHIS sampling design are currently available only for the software packages WesVar, SUDAAN, SAS, and STATA.

The opinions expressed here are those of the author. The material is being presented on behalf of CAFA; none of it has been reviewed or approved by the staff of the UCLA Center for Health Policy Research, the California Department of Health Services, or the Public Health Institute.