# SHADAC

# Issue Brief

**A Health Data Resource for States**

## Overview of Approaches for Estimating Uninsurance Rates at the Sub-state Level

Many states now conduct state household surveys to estimate health insurance coverage and some states have begun exploring methods to derive coverage estimates for different populations, specifically for geographic areas (regions, counties) and racial/ethnic sub-populations within their borders. The purpose of this issue brief is to highlight three approaches that have been used to estimate uninsurance rates at the sub-state level. We provide an overview of the conceptual and methodological issues involved in estimating uninsurance rates at the sub-state level, assess the relative strengths and weaknesses of each approach and conclude with a list of resources useful to readers interested in learning more about small-area estimation.

### DIRECT APPROACH THROUGH SURVEY DESIGN AND SAMPLING

The direct approach to estimating health insurance coverage within a small area (e.g., a county or city) can be characterized by two features: (1) Use of a measurement instrument (e.g. state survey) to directly measure health insurance coverage, and (2) measurements from a sample of people drawn from the actual population of interest (e.g., the county or city of interest). For example, to directly measure health insurance coverage within a specific county, researchers could construct a survey instrument designed to measure health insurance and draw a sample of people from the county to serve as survey respondents.

Three conditions need to be met in order to obtain high-quality direct estimates of health insurance coverage. First, the instrument used to measure the concept should be valid. For an instrument to be valid the survey items need to do a good job of determining whether or not someone has health insurance coverage. Second, each member of the population of interest should have a known probability of selection into the sample. For example, if you are conducting a survey of 500 people and you draw a simple random sample from a population list that includes all 5000 people in the county, then each person's probability of selection would be 10 percent. The final condition that needs to be met in order to derive direct estimates is to have a large enough sample size. A good rule of thumb is that the equivalent of 100 simple random sample cases are needed for each population of interest.

Although direct estimates provide the most defensible estimates, they are also the most costly to produce. Indeed, the cost of producing high-quality direct estimates for small areas is often prohibitive. When at least one of the three conditions to derive direct estimates is not met, people often turn to one or more other approaches, depending on their expertise, resources, and data available. The simplest of these alternatives is the "proxy measure" approach to small area estimation.

## PROXY MEASURE APPROACH

The proxy measure approach uses some measure that can serve as a proxy of health insurance coverage to estimate health insurance coverage, and that proxy measure is generally applied to a proxy population within a county. A commonly used proxy measure of uninsurance uses administrative records from all the hospitals within a county to determine the percent of specified discharge diagnoses that were coded as "self-pay." Specifically, this would entail extracting information on the *expected* primary source of reimbursement reported on in hospital discharge data sets from all hospitals in an area for specifically chosen diagnoses. Patients discharged with one of these specific diagnoses who are classified as "self-pay" (meaning the person, and not an insurance company or the government, was expected to pay the bill) would be designated as being uninsured. For example, if 8 percent of all patients with these diagnoses were expected to self-pay, then the uninsurance rate in the county could be set at 8 percent as well.

A major strength of this and other proxy measures is the low cost. These data are relatively inexpensive to compile and are routinely collected in a majority of states. Moreover, the use of this particular proxy measure avoids the problem of basing estimates on small survey samples since generally there will be reasonably large numbers of discharges for the selected diagnoses within a specific geographic area.

There are some concerns with bias and measurement error. Not everyone who was discharged from each hospital is going to be a resident of that county, which can bias the estimate for the referent county. And a given county's estimated rate of uninsurance can also be biased from its *actual* rate because not every patient living in the county will have gone to one of the county's hospitals.

Furthermore, for the diagnoses selected for use in this analysis it is critically important that the decision to be admitted to a hospital be completely *independent* of whether one has insurance coverage or not. For example, for a given diagnosis with some 'discretion' about the need to be hospitalized, individuals with insurance coverage are more likely to be admitted to a hospital

than those without coverage. To the extent that you include this type of diagnosis in the set of diagnoses forming your overall proxy measure of uninsurance, you would underestimate the amount of uninsurance in the county. Although this "self-pay" proxy measure uses data from the county of interest, it is nevertheless a "proxy" population that can be expected to yield an estimate of uninsurance of greater or lesser accuracy.

Finally, because *actual* insurance coverage is only correlated with expected self-pay and is not the same thing, use of this proxy measure of coverage can involve error. For example, an individual may be classified as "self-pay" at the time of discharge but receive retroactive Medicaid coverage for this hospital expense later. As this example shows, using this proxy measure would yield too high an estimated uninsurance rate unless some adjustment to it could be made to account for this kind of error. This type of adjustment is difficult to do—and subject to imprecision—with only expected primary payment data available.

Although proxy measures often have fairly large sample sizes (for example expected payer information on discharges from all hospitals within a county for an entire year), the proxy measures approach is generally considered a last resort. With proxy measures the potential for bias is high. If there is nothing else available, you may want to consider it. At a minimum, however, you should exercise great care in selecting the proxy used, preferably using only those that have been rigorously evaluated for potential bias.

## MODEL-BASED APPROACH

When the sample size within a geographic area is too small, or there are no national or state survey data on insurance coverage available, the previously described direct estimation is not possible or desirable. Under these conditions, statisticians and researchers must use several sources of data and statistical analyses to develop direct and indirect estimates of health insurance coverage. We illustrate the spectrum of model-based approaches with a "simple model-based approach" and a "complex model-based approach."

## Simple Model-based Approach

The simple modeling approach predicts health insurance coverage for a specific geographic area using, 1) one or more variables correlated with health insurance coverage and, 2) correlation based on data obtained from the geographic area of interest. It then is possible to predict coverage for other geographic areas that do not have a measure of health insurance coverage by inserting the values of the correlated measures into the models and use this model-based estimate as the health insurance coverage estimate.

An example of this approach is using unemployment rates to estimate the level of uninsurance. The use of unemployment rates is attractive for two reasons: 1) unemployment rates are correlated with health insurance coverage rates, and 2) unemployment rates are available for every county in the United States from the Bureau of Labor Statistics and in a timely manner.

If, for example, it was found through statistical analysis that the uninsurance rate was, on average, 1.5 times the amount of the unemployment rates across a large number of counties, then in counties, without any direct measure of uninsurance, an estimate of uninsurance would be 1.5 times the unemployment rate prevailing in the county. With such a simple model it is clearly preferable that the counties used to develop the model be as demographically similar as possible, be located within the same state, and be as close as possible to the counties using the model to predict their uninsurance rates.

## Complex Model-based Approach

The pre-eminent example of this model-based approach in current use—unfortunately not for uninsurance—is the Census Bureau's Small-Area Income and Poverty Estimates program (known by it acronym SAIPE). In the SAIPE program, up-to-date estimates of the number of school-age children living in poverty in U.S. counties are obtained from a combination of two estimates. First, and for those counties that have been sampled by the annual March Supplement to the Current Population Survey (CPS), this survey provides a direct estimate of the number of school-aged children in poverty. Even for counties that have been sampled, however, this direct estimate is usually based on very small samples. As a result, even if three years of March CPS information are combined to form one direct estimate, it is still likely to be subject to too large an amount of sampling error to be of much policy utility if used alone. In addition, only about one-third of counties nation-wide are included in the March CPS sample in any given year, and consequently no direct estimate is possible for the majority of counties in the country.

To overcome this deficiency, researchers have developed regression models to provide indirect, or synthetic, estimates of a county's number of school-age children in poverty. This approach begins by assembling a large data set on all the counties in the entire country that have been included in the CPS samples. The data collected for this project come from the CPS itself, on each county's number of school-age children in poverty, plus Internal Revenue Service (IRS) data on individual tax returns and data from the federal food stamp program, all aggregated to the county level to yield predictors of school-age children in poverty. That is, these predictors include such county-specific measures as the number of child exemptions reported by families in poverty in the county, and the number of people receiving food stamps in the county. These data are then used in regression models to establish the statistical relationship between the expected number of school-age children in poverty in each county and the levels of these predictor variables for the county. Importantly, these predictor variables are selected in part because of the feasibility (for the Census Bureau) of obtaining reasonably up-to-date values for them for all the counties in the country. Thus it is possible to use these up-to-date predictor values to estimate each county's number of school-age children in poverty. Finally, since this regression model has been estimated on a large data set (all counties in the county with CPS samples), the synthetic or indirect estimates derived from it are capable of achieving reasonably high levels of 'predictive' accuracy.

The SAIPE model estimates of school-age children in poverty are formed as a mixture of the direct estimates (for counties included in the March CPS sample) and the model predictions, or indirect estimates. By blending these two estimates together in a sophisticated manner that takes into account the accuracy of each estimate, the resulting blended estimate is better than either direct or model-based estimate would be alone. Importantly, they also provide an estimate for those counties not included in the March CPS samples. The other advantages of the SAIPE model estimates are that they can be updated on an annual or biennial schedule; and they can be expected to have less error than using outdated census estimates, the alternative to them. The major disadvantage is that the production of these model-based estimates requires substantial resources. These models must be developed initially and then
evaluated by highly-trained statisticians; they require access to large amounts of data, preferably nationwide, all of which may not be in the public domain; and the models themselves must be updated periodically, which also entails large resource costs.

## DISCUSSION/CONCLUSION

Desirable levels of accuracy for well-defined sub-populations and specific time periods at the sub-state level are obtainable only at very substantial costs, since they are achievable only from large-sample based direct estimates. Conversely, estimates using proxy measures are generally possible with low resource costs but are very unlikely to provide sufficient accuracy or sensitivity to be useful for most evaluation purposes. Specifically, the proxy measure and model-based approaches in general will not be sensitive to specific interventions within a geographic area. For example, if a county implements an intervention to increase insurance coverage, it's impact will only be detectable from a model if either:
(1) one or more of the correlates are directly impacted by the intervention itself and hence are directly related to uninsurance status (e.g. "self-pay" status for specific diagnoses); or (2) there is a significant number of directly measured cases from the area in the blended-model (in which case it

really becomes best to use the direct estimate approach). Thus, complex, difficult to achieve and/or costly requirements are placed on measure proxy and blended-model approaches if they are to serve the needs of most evaluation uses.

Model-based estimates could prove considerably more useful, were a counterpart to the SAIPE model estimates for children in poverty ever developed by the Census Bureau for uninsurance in small-areas, producing what might be called Small-Area Uninsurance Rate Estimates (SAURE). They would be based on a large data set, again including all the counties in the country with a sample in the March Supplement to the CPS. And they could use many predictor variables available only to the Census Bureau and on a reasonably timely basis. These models are capable of generating estimates with reasonably high predictive accuracy and in a reasonably timely manner. But like the SAIPE model estimates for children in poverty, these Small-Area Uninsurance Rate Estimates (SAURE) would have to be a three-year average estimate. And this three-year time dimension would not accommodate many evaluation uses, although it might prove satisfactory for less rigorous monitoring purposes. Nonetheless, SHADAC is working with staff at the Census Bureau to assess the feasibility of estimating uninsurance rates in small areas using the CPS.

In conclusion, selection of the appropriate estimation approach is not straightforward and requires an assessment of the principal strengths and weaknesses of each approach (Table 1). Unfortunately, each of the previously listed desired properties for small-area estimates of uninsurance is achievable only at the price of steep trade-offs among the others. When evaluating the relative merits of the various approaches described, one must also consider the ease or unease with which the results can be described. Specifically, it will be important (and difficult) to provide policymakers with an appropriate understanding of the complex statistical and methodological issues associated with the proxy direct and model-based approaches. End users of the information generated by the approaches must also be informed of the requisite cautions to guard against over-interpretation of the data.

**TABLE 1. SUMMARY OF PRINCIPAL STRENGTHS AND WEAKNESSES OF THE DIFFERENT APPROACHES TO ESTIMATE UNINSURANCE RATES AT THE SUB-STATE LEVEL**

| APPROACH | PRINCIPAL STRENGTH | PRINCIPAL WEAKNESS |
|---|---|---|
| Direct Estimation Through Survey Sampling | Precision | Cost |
| Proxy Measure Approach | Cost | Bias |
| Model-based Approach | Predictive accuracy | Complexity |

# Related References

### USE OF HOSPITAL ADMINISTRATIVE DATA

Rask, K.J. (1994). Hospital discharge data and the uninsured. Journal of Health Care for the Poor and Underserved, 5, 275-279.

Turner, C., & Campbell, E. (1999). Counting the uninsured using state-level hospitalization data. Public Health Reports, 114, 149-156.

### MODEL-BASED APPROACHES

Chand, N., & Alexander, C.H. (1999). Using administrative records for small area estimation in the American Community Survey. http://www.fcsm.gov/papers/mcf.html

Chand, N., & Malec, D. (2001). Small area estimates from the American Community Survey using a housing unit model. http://www.fcsm.gov/01_papers/Chand.pdf

Diehr, P., Madden, C.W., Cheadle, A., Patrick, D., Fishman, P., Char, P., & Skillman, S. (1991). Estimating county percentages of people without health insurance. Inquiry, 28, 413-419.

Malec, D., Davis, W.W., & Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. Statistics in Medicine, 18, 189-200.

Malec, D., Sedransk, J., Moriarty, C., & LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. Journal of the American Statistical Association, 92, 825-826.

National Research Council (2000). Small area income and poverty estimates: Priorities for 2000 and beyond. Washington, D.C., Committee on National Statistics, National Academy of Sciences.

Popoff, C., Judson, D.H., & Fadali, B. (2001). Measuring the number of people without health insurance: A test of a synthetic estimates approach for small areas using SIPP microdata. http://www.fcsm.gov/01_papers/Popoff.pdf

Schaible, W.L., Brock, D.B., Casady, R.J., & Schnack, G.A. (1979). Small area estimation: An empirical comparison of conventional and synthetic estimators for states. U.S. Dept. of Health, Education, and Welfare, Public Health Service, Office of Health Research, Statistics, and Technology, National Center for Health Statistics; ISBN: 084060176X.

Schirm, A.L., Zaslavsky, A.M., & Czajka, J.L. (1999). Large numbers of estimates for small areas. http://www.fcsm.gov/papers/schirm.pdf