CDC initiatives for the NBCCEDP: Design, Testing and Implementation of Small Area Estimation of the Number of NBCCEDP Eligible Women.

Initial Assessment of Small Area Estimation of the Number of Eligible Women for the CDC's NBCCEDP

September 2006

Brett O'Hara
Joanna Turner
Mark Bauder
Steven Riesz
David Waddington

Submitted to:
Centers for Disease Control and Prevention
Division of Cancer Prevention and Control
4770 Buford Highway NE, MS K-55
Atlanta, GA 30341-3717

Technical Monitor:
Florence K. Tangka

Submitted by:
U.S. Census Bureau
Small Area Estimates Branch
Room 1451 Building 3
Washington, DC 20233-8500
(301) 763-3193

Task Manager:
Joanna Turner

Project Director:
David Waddington

ACKNOWLEDGMENTS

This assessment of using small area techniques for estimating low-income uninsured women has benefited greatly from the advice and support of many individuals. Florence Tangka, the Centers for Disease Control and Prevention's (CDC's) project technical monitor, provided guidance throughout the process. James Gardner and Janet Royalty facilitated data exchange and increased our understanding of the programmatic needs. While visiting the CDC, we received help from numerous staff and management to improve our estimates.

The project team also benefited substantially from colleagues. Joseph Dalaker provided the initial data runs on the direct estimates. Lucinda Dalzell negotiated for the data that was used in the model. Colleen Joyce provided useful advice on cartographic issues. David Powers helped in the quality control process of the data. Finally, Kirk Davis provided programming support.

David Waddington
U.S. Census Bureau
September 2006

CONTENTS

Chapter **Page**

Chapter **Page**

SUPPORTING MATERIALS

Fisher, Robin and Bauder, Mark (2006). "A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html

Fisher, Robin and Riesz, Steven (2006). "A Model for the State-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html

EXECUTIVE SUMMARY

The Centers for Disease Control and Prevention (CDC) have a congressional mandate to provide screening services for breast and cervical cancer to low-income, uninsured, and underserved women through the National Breast and Cervical Cancer Early Detection Program (NBCCEDP). Currently, the NBCCEDP produces national and state participation rates for this program. Participation rates serve many functions such as evaluation of long-term performance by states and as an aid to grantees in prioritizing their resources toward at risk groups. Currently, calculating these participation rates relies on direct estimates from the Annual Social and Economic Supplement of the Current Population Survey, computed and tabulated by the U.S. Census Bureau. The state estimates contain 2-year and 3-year averages of the number of uninsured low-income women by age group (18-64, 40-64, and 50-64). The national estimates include the same information by race and ethnicity.

The purpose of this project is to investigate the feasibility of estimating the numbers of eligible women by race and ethnicity, age, and poverty level (0-200 and 0-250 percent of the federal poverty threshold) for states and counties. Because of small sample sizes, direct estimates alone cannot be used for these sub-groups. These estimates are required for calculating the screening rates for most of the NBCCEDP grantees. This report is a product of the second phase of our cooperative work. Phase one dealt primarily with data acquisition and model development for the uninsured. Phase two assesses the feasibility of producing estimates of the number of low-income uninsured women by age. Additionally, we investigated race and ethnicity categories. This work builds on the Census Bureau's Small Area Health Insurance Estimates (SAHIE) program, which provided county and state health insurance coverage estimates for 2000 by age.

We conducted an extensive review of the literature on health insurance to determine necessary information for the model. When possible, we collected the relevant data. Otherwise, we are in the process of obtaining the needed data. Model development began for county and state small area estimates (SAE) of women eligible for the NBCCEDP, by income to poverty ratios (IPRs), age, race, and ethnicity. For the feasibility study, the IPRs and age categories are the same as already mentioned. The race and ethnicity categories are currently non-Hispanic Black, non-Hispanic non-Black, and Hispanic. After the feasibility study, the race and ethnicity categories will be changed to reflect standard categories, as discussed with CDC staff in October 2005.

We have established the feasibility for producing estimates for the states and counties. The current version of our model of the NBCCEDP eligible population has produced estimates by age, race and ethnicity, and IPR for states. We have also produced estimates by age and IPR for counties, but some counties may need to be aggregated into larger areas. However, many simplifying assumptions have been made. In phase three, these assumptions will be examined, new categories of race and ethnicity will be incorporated, and new data will be added. We expect the precision of the estimates to improve. It remains to be decided if the current level of precision is adequate for CDC's purpose.

TECHNICAL FINDINGS

The technical findings are both specific and general for improving the model and its usefulness in producing estimates of the number of women eligible for the NBCCEDP.

- We can model age, race and ethnicity, and income to poverty ratios (IPRs) at the state-level.
    - Race and ethnicity will be changed to reflect the Census Bureau's standards: non-Hispanic White, non-Hispanic Black, and Hispanic. The remainder of non-Hispanics is grouped as non-Hispanic Other.
    - Estimates for a non-Hispanic Other category are likely to have low reliability due to heterogeneity.
    - We can produce the appropriate IPR estimate ($\leq$ 200 percent or $\leq$ 250 percent of the federal poverty threshold) based on the states eligibility requirements.
    - The current state-level model should be treated as a research product; there are reasons to believe the model form is superior to the county-level model but implementation is more difficult.
- We can model eligibles by age and IPR at the county-level. The feasibility of modeling other dimensions is unknown.
    - If the reliability of an estimate is insufficient, the Census Bureau's policy is to aggregate to larger areas or broader categories.
    - It remains to be decided if the current level of precision is adequate for CDC's purpose, particularly for women ages 50-64.
- We need to make state and county model improvements before we can publish estimates on the Census Bureau web site.
    - Variance models do not fit well when proportions are close to zero or 100 percent within an age and sex category. This is particularly true for low-income women ages 50-64.
    - The important assumption of normality in the models needs to be tested during phase three.
    - The important assumption of conditional independence in the county-level models needs to be tested during phase three. If the assumption is imprecise, re-specifying the county-level model will take significant effort. Preliminary results indicate that this will not be necessary.
- We need to expand the predictive variables for income categories and health insurance coverage for smaller coefficients of variation (CVs).
    - The average CV for uninsured women ages 40-64 $\leq$ 200 percent of the federal poverty threshold in the current county model is 0.36.
    - The average CV for uninsured women ages 40-64 $\leq$ 200 percent of the federal poverty threshold in the current state model is 0.22.
- We need to create participation rates with respect to relevant income categories by state.

In the second phase of this work, we evaluated the feasibility of producing estimates of low-income uninsured women by age. Additionally, we investigated the feasibility of producing estimates by race and ethnicity at the state level. The following is a summation of estimates that, based on current work, we believe feasible.

Table 1: Low-income* uninsured women by state and county#

	18-64		40-64		50-64	
	State	County	State	County	State	County
All low-income uninsured women by age	Y	Y	Y	Y	Y	M□
Non-Hispanic Non-Black^	Y	-	Y	-	Y	-
Non-Hispanic Black	Y	U	Y	U	Y	U
Hispanic	Y	U	Y	U	Y	U
Non-Hispanic White	M	U	M	U	M	U
Non-Hispanic Other	N	N	N	N	N	N
Non-Hispanic by the following race categories: American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander	N	N	N	N	N	N

Y: We have produced this estimate.
M: We think we can produce this estimate.
U: It is unlikely that we can produce this estimate, but we will look further into the issue.
N: It is not feasible to produce this estimate.

* Each state program defines low-income as 200 percent, 250 percent, or (for Nebraska) 225 percent of the poverty threshold. The SAHIE program will provide the relevant poverty threshold for each state. For Nebraska, the CDC advises using 200 percent of the poverty threshold.

It is not feasible to produce estimates for tribal areas.

□ Our confidence intervals are currently too large to make reliable estimates.

^ The race and ethnicity categories for the feasibility study are non-Hispanic Black, non-Hispanic non-Black, and Hispanic. After the feasibility study, the race and ethnicity categories will be changed to reflect standard categories, as discussed with CDC staff in October 2005. These standard categories are non-Hispanic White, non-Hispanic Black, Hispanic, and non-Hispanic Other.

OVERVIEW OF DOCUMENTATION

The feasibility study includes twelve chapters discussing the background of this project, the model design, available data, and Census Bureau standards. Appendix A is a table of acronyms.

The county estimates have not been adjusted to add up to the state estimates. The following provisional estimates and coefficients of variation (CVs), as well as accompanying maps, are available for the Centers for Disease Control and Prevention (CDC) to review:

- The number of uninsured women ages 18-64, 40-64, and 50-64 ≤ 200 percent of the federal poverty threshold at the state-level and CVs of the estimates.
- The number of uninsured women ages 18-64 and 40-64 ≤ 200 percent of the federal poverty threshold at the county-level and CVs of the estimates. In next years deliverable, we will know if ages 50-64 will be reliable.

These estimates are experimental and are provided using the Census Bureau's pre-release agreement. They are for the CDC's internal use to help determine if the estimates are adequate for its purposes.

Technical documents on the methods used to construct our estimates are available on the SAHIE program's web site.[1] More information on the CVs is available in these documents.

[1] http://www.census.gov/hhes/www/sahie/publications.html.

PART ONE

BACKGROUND

I. National Breast and Cervical Cancer Early Detection Program[2]

The Breast and Cervical Cancer Mortality Prevention Act of 1990 established the National Breast and Cervical Cancer Early Detection Program (NBCCEDP) in 1991. This act authorizes the Centers for Disease Control and Prevention (CDC) to provide critical breast and cervical cancer screening services to underserved or uninsured women. In particular, women ages 50-64, women with low incomes and women from minority groups are priority groups for receiving services. The CDC collaborates with state-based grantees to develop comprehensive screening methods and best practice guidelines. Because the NBCCEDP operates under a series of cooperative agreements with the grantees, the NBCCEDP programs vary by state. For instance, states may direct their efforts based on state-specific mortality trends or known high risk sub-populations and these efforts are carried out using different prevention strategies and organizational structures. The NBCCEDP grantees collect surveillance data on women served through the program and report data elements to the CDC that describe demographic characteristics, screening history, and screening and diagnostic outcomes for these women. The NBCCEDP under the Act of 1990 could only provide funds for screening and diagnostic services while the state programs had to secure other resources for treatment.

The Breast and Cervical Cancer Prevention and Treatment Act of 2000 gave states the option to use Medicaid funding to provide medical assistance to eligible non-elderly women who were screened by the NBCCEDP and required treatment. The enhanced federal match rate (the percent of federal expenditures for Medicaid services) under this act varies between 65-84 percent with the remainder paid by the state. The match rate is based on a formula that factors in the number of low-income uninsured children and the annual wages in the health care industry.

The Native American Breast and Cervical Cancer Treatment Technical Amendment Act of 2001 clarified the meaning of health insurance coverage by stating that Indian Health Services or other Indian health programs do not count as health insurance coverage for purposes of the NBCCEDP. To be eligible for Medicaid coverage under the Act of 2000/01, a non-elderly woman: 1) cannot be eligible for regular Medicaid, 2) cannot have comprehensive health insurance, 3) must be screened through the NBCCEDP, and 4) must need treatment for breast and/or cervical cancer. All 50 states, the District of Columbia, 4 territories, and 13 American Indian and Alaska Native organizations have received approved Medicaid options for treating NBCCEDP identified patients.

These laws were passed to provide critical services for uninsured, low-income women at risk of breast or cervical cancer. Studies showed that early intervention was cost effective and saved lives. The NBCCEDP program was funded for $210 million in fiscal year 2004 to provide both screening and diagnostic services. These services include: clinical breast examinations, mammograms, Pap tests, surgical consultations, and diagnostic testing for women whose screening outcome is abnormal. Since inception,

[2] The information from this section was gathered from http://www.cdc.gov/cancer/nbccedp.

the program has screened nearly 2 million women and diagnosed 17,009 breast cancers, 61,474 precancerous cervical lesions, and 1,157 cervical cancers.

Federal legislation enacting the NBCCEDP mandates that women below the federal poverty level must have access to program services free of charge. Ostensibly, services may be provided to other low-income women as long as free services remain available to those women below the federal poverty level. Typically, the states chose 200 or 250 percent of the federal poverty level to define "low-income" for women to be eligible for the program. Many of the state-based programs distribute the funds to give preferential treatment to older women.

Under normal circumstances, medical guidelines suggest that a woman should have a screening mammogram every 1 to 2 years from ages 40 to 74.[3] For cervical cancer screening, the first screening should be within 3 years of onset of sexual activity or age 21 and follow-up screening at least every 3 years.[4] For both breast and invasive cervical cancer, the highest risk category is from ages 50-64. Seventy-five percent of all diagnosed cases of breast cancer are among women age 50 years or older.[5]

As emphasized on the NBCCEDP web site concerning the outcomes of the state-based programs, the de facto eligibility requirements vary from state to state depending on the programs' emphasis on age, race, or ethnicity. "Comparing data across grantee programs is not advised due to extensive variation across programs."[6]

[3] http://www.cdc.gov/cancer/nbccedp/info-bc.htm
[4] http://www.cdc.gov/std/HPV/ScreeningTables.pdf
[5] http://www.cdc.gov/cancer/nbccedp/about2004.htm
[6] http://www.cdc.gov/cancer/nbccedp/sps/profiles/national_aggregate.htm

II. Need for Estimates of Eligibles

The CDC is interested in estimates of the number of women eligible for the NBCCEDP and of those eligible what percentage are being screened through the program. Participation rates serve many functions.

- Improve the analysis and use of information gathered systematically from grantees.
- Assess the supply and demand for cancer screening services at different levels of geography and among priority populations.
- Identify the determinants of and disparities in use of screening, diagnostic and preventive services, and treatment for cancer.
- Evaluate the geographic distribution of the NBCCEDP eligible women.
- Prioritize outreach strategies to reach women who rarely or never have been screened. Results can be useful to inform studies on issues of outreach and capacity within state programs.
- Develop resource allocation strategies not previously available to most programs and provide CDC monitors ability to track screening progress below the state level.
- Measure the success in reaching specific groups such as Hispanics.
- Measure long-term performance of grantees.

Currently, the CDC has a variety of participation rates at the national level.[7]

- 7–10 percent of U.S. women of screening age are eligible to receive NBCCEDP services.
- 7 percent of program-eligible women and 1 percent of all U.S. women ages 18-64 received NBCCEDP services for cervical cancer.
- 13 percent of program-eligible women and 1 percent of all U.S. women ages 40-64 received NBCCEDP services for breast cancer.
- 20–21 percent of eligible women ages 50-64 received Pap tests and/or mammograms through NBCCEDP.

To estimate detailed participation rates, the CDC requires estimates of the numbers of uninsured women by race and ethnicity, age, and income for the states and counties in the United States. These numbers will improve the calculation of the participation rate. The age, race, ethnicity, and income details are particularly important because of the state-to-state and county-to-county variations in the programs. If the priority group for a state is low-income uninsured Hispanic women, the number of women eligible should reflect this information. The definition of low income also varies by states, typically 200 or 250 percent of the poverty threshold. To the extent possible, the CDC wants the number of program-eligible women by race and ethnicity, age categories (18-64, 40-64, and 50-64), and income to poverty ratios (IPRs) (0-200 and 0-250 percent of the federal poverty level) at the county and state levels. These categories correspond to the majority of women served by the state-specific NBCCEDP.

[7] See http://www.cdc.gov/cancer/nbccedp/sps/ and http://www.cdc.gov/cancer/nbccedp/about2004.htm

The number of eligibles for the NBCCEDP cannot be calculated by direct survey estimates. At the state-level, direct estimates of uninsured women with low-income have large variances. When the number of women is further parsed into age, race, and ethnicity categories, some categories at the state-level cannot be estimated by direct survey estimates. Direct estimates of health insurance coverage do not exist for most counties and the aforementioned problems are exacerbated.

III. U.S. Census Bureau's Small Area Health Insurance Estimates Program[8]

The CDC project is building on previous work of the U.S. Census Bureau. The Small Area Health Insurance Estimates (SAHIE) program was created to develop model-based estimates of health insurance coverage by age for counties and states. County-level data on health insurance coverage are not available elsewhere because neither the decennial census nor the American Community Survey contain questions on this topic. National surveys such as the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) do not have sufficient sample to provide direct estimates at the county level. The SAHIE program's focus is to produce a consistent set of nation-wide estimates for sub-state areas.

The SAHIE program has developed model-based estimates for calendar year 2000 by counties and states. These estimates include:

- total population with and without health insurance coverage;
- children under age 18 with and without health insurance coverage; and
- measures of uncertainty of the estimates.

We model county-level health insurance coverage by combining survey data with population estimates and administrative records. Our estimates are based on data from the following sources:

- CPS ASEC;
- demographic population estimates;
- aggregated federal tax returns;
- food stamp participation records; and
- Medicaid participation records.

The SAHIE program's state-level estimates are aggregated from the county estimates. The estimates are adjusted so that, before rounding, county numbers sum to their states and similarly the states sum to the CPS ASEC national estimates.

This is a new program at the Census Bureau and the first ever set of estimates were released in July 2005. The Small Area Estimates Branch (SAEB) supports two programs, the Small Area Income and Poverty Estimates (SAIPE) program and the SAHIE program. The SAHIE program builds on the work of the SAIPE program. For nearly ten years the SAIPE program has been producing model-based estimates of poverty and median household income for counties and states, and estimates of poverty and population for school districts. This program has been favorably reviewed by the National Acadamy of Sciences and other organizations. These estimates are used for the administration of federal programs and the allocation of federal funds to local jurisdictions, such as under Title I of the No Child Left Behind Act of 2001.

[8] More information is available at http://www.census.gov/hhes/www/sahie.

The SAHIE program's methodology and estimates have undergone internal Census Bureau review as well as external review. Our preliminary research was evaluated favorably by the 2002 Census Advisory Committee of Professional Associations. In February-April of 2005, representatives from other federal agencies, the State Data Centers (SDCs), the Federal and State Cooperative Program for Population Estimates (FSCPE), and the University of Minnesota's State Health Access Data Assistance Center (SHADAC) participated in a formal evaluation of our methodology and experimental estimates.

The feedback was useful in directing our attention to specific issues and informing our thinking generally. Based on the comments received, we have made improvements to the methodology and have plans for further research. Based on availability of resources, we are considering making enhancements to our estimation methods and producing estimates for additional years and age groups.

PART TWO

DESIGN FOR ESTIMATION

IV. Estimation Strategy

The first part of the Centers for Disease Control and Prevention (CDC) project is to demonstrate that useful estimates can be made. The Small Area Health Insurance Estimates (SAHIE) team developed a model that estimates the number of insured persons. We estimate the number of insured persons for counties by age, sex, income categories and, for states, by age, sex, income categories, and race and ethnicity. We fit models using a collection of covariates deemed useful by previous work[9] and examine the models to see how the covariates fit and whether the modeling assumptions are violated. After a model is fit successfully, we will examine measures of variability to decide whether the estimates are suitable for use.

There are differences between the models we use for state and county-level estimation. These differences are primarily due to issues of computation time.

Details for the models are presented in the supporting materials;[10] the models are described more generally in the following sections.

A. Hierarchical Bayesian Model

The models for both state and county level estimates are hierarchical models, estimated using Bayesian methods. These methods require numerical algorithms for all but the simplest models, and we use Markov Chain Monte Carlo (MCMC) techniques.[11] This allows for a flexible class of models. Any model for which the conditional distributions can be calculated or approximated in computer code can, in principle, be estimated this way. Quantities which must be approximated in non-Bayesian methods, and for which the approximations are only known for certain sets of modeling assumptions, such as some standard errors,[12] can be calculated almost exactly in the MCMC. The 'almost' arises from the fact that the Monte Carlo algorithm itself contributes some random error, which approaches zero as the number of iterations in the algorithm gets large.

The following sections give more detail about the methods we use in this study, including the general forms of the models, the implementation of the models, problems we have encountered, and methods for assessing results.

---

[9] Fisher, Robin and Turner, Joanna (2004). "Small Area Estimation of Health Insurance Coverage From the Current Population Survey's Annual Social and Economic Supplement and the Survey of Income and Program Participation", *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: The American Statistical Association.

[10] Fisher, Robin and Bauder, Mark (2006). "A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html.
Fisher, Robin and Riesz, Steven (2006). "A Model for the State-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html.

[11] Gilks, W.R. *et al*. eds (1998). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

[12] Prasad, N.G.N. and Rao, J.N.K (1990). "The Estimation of the Mean Squared Error of Small-Area Estimators", *Journal of the American Statistical Association*, 85, 163-171.

B. General Details of Model

This section provides a basic discussion of the commonalities and differences between the state and county models. Details on the state and county models are provided in the next two chapters.

Ideally, we would produce estimates for the number of uninsured low-income women by age, race, Hispanic origin, and income to poverty ratio (IPR) categories at the county and state level. For this study, we use five age categories (0-17, 18-39, 40-49, 50-64, 65 and older), three race and Hispanic origin categories (non-Hispanic Black, non-Hispanic non-Black, Hispanic), two sex categories, and three IPR categories (0-200, 200-250, and above 250 percent of the federal poverty threshold). For statistical purposes, the entire population was used in the model including all age groups and both sexes.

Because of smaller sample sizes in counties, we produce estimates for fewer categories at the county level than at the state level. For states, we estimate the number uninsured by IPR, age, race, sex, and Hispanic origin (ARSH) categories. For counties, we estimate the number uninsured by IPR, age and sex, but not by race and ethnicity categories. [13]

A small area estimate is often referred to as a *domain*. If the estimate of interest were the number of uninsured persons by state, there would be one domain per state and a total of 51 domains. As the number of domains increase, the estimation problem becomes more difficult. The state model has 90 domains for each state and the county model has 30 domains for each county. Alternatively stated, the state model estimates a total of 4,590 domains and the county model estimates a total of 94,260 domains.

We estimate the number insured within an IPR category for a given demographic group by estimating two proportions. One is the proportion of those in the demographic group who are in the IPR category. The second is the proportion of those in the IPR category within the demographic group who are insured.

1. Predictors for the IPR and Insurance Coverage Estimates

IPR estimate: The predictors for the number of people in each IPR category include:

- indicators of sex and age categories (and race and ethnicity in the state model);
- proportion of Internal Revenue Service (IRS) exemptions in an age group that are in the IPR category;
- proportion in an IPR category according to the decennial census;
- proportion receiving food stamps in the state or county;
- proportion participating in Medicaid within an age group in the state or county;

[13] We will further investigate whether the sample size for the counties is an impediment to increasing the number of domains in the county model and we will be changing the race categories as described in the technical findings.

- mean log aggregate gross income (from tax records);
- variance of the log aggregate gross income within the state or county; and
- indicator of being in the South Census region.

Note that for the predictors based on tax records, there are two age groups defined by child and adult exemptions (assumed to be under 18 and 18 and over) and two IPR categories for the two age groups (less than or equal to 200 percent of poverty and over 200 percent of poverty).

Insurance Coverage estimate: The predictors for proportion insured in a domain include:

- indicators of the domain;
- indicators of being in the South Census region, and of being in the West Census region;
- proportion getting food stamps in the county interacting with the IPR category, so there is a different coefficient in each IPR category;
- proportion participating in Medicaid in age category interacting with the IPR category, so there is a different coefficient for each IPR category;
- mean log aggregate gross income (from tax records); and
- variance of the log aggregate gross income within the state or county.

Note that for both IPR and insurance coverage, the current models for state-level estimates do not use all of the above predictors. Note also that in some cases, the models use transformations of the predictors. We consider the current set of predictors to be preliminary.

2. Differences between the State and County Models

The state model draws from the tradition of the directed acyclic graph (DAG) models.[14] The unobserved proportions, or a transformation of the proportions, are modeled as drawn from some distribution. The dependencies of the various data on these proportions are also modeled.

The county-level model is a variation of a generalized linear model (GLIM) with random effects, often seen in small area estimation problems.[15] These models have the advantage

[14] The DAG literature is extensive. Here is a seminal paper. Lauritzen, L. and Wermuth, N. (1983). "Graphical and Recursive Models for Contingency Tables", *Biometrika*, 70, 537-552.

[15] Examples include: **Fay, R. and Herriot, R. (1979).** "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", *Journal of the American Statisitcal Association*, 85, 398-409. **Fisher, R. (1997).** "Methods Used for Small Area Poverty and Income Estimation", *Proceedings of the Joint Statistical Meetings, Section on Governmental and Social Statistics*, Alexandria, VA: The American Statistical Association. **Ghosh, M., Natarajan, K., Stroud, T.W.F., and Carlin, B. (1998)**. "Generalized Linear Models for Small Area-Estimation", *Journal of the American Statistical Association*, 93, 273-282. **National Research Council (2000*).*** *Small Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*, Panel on Estimates of Poverty for Small Geographic Areas, Constance F. Citro and Graham Kalton, editors. Committee on National Statistics. Washington DC: National Academy Press.

that people are familiar with them and it may be that the extra flexibility of the DAG models is not necessary.

For states, the IPR and insurance coverage estimates are modeled jointly. There are about 20 times as many domains for the county-level estimates as for the state-level estimates. Because of the larger number of domains, we found that it would take an infeasible amount of computer time to implement the same model for counties as for states. Because of these computational issues, we model the numbers in an IPR and the proportion insured independently.

We implemented the state model in WinBUGS. Even with the differences in the county and state models, standardized software is not feasible for the county model because of the length of the run times. Therefore, the county model was implemented in C.

V. State Model

A. Model

Bayesian methods were used to calculate posterior means and standard deviations. Bayesian models are not tractable with analytical approaches, so we use MCMC algorithms. These algorithms have several advantages. They allow estimation in a very large class of models; even large changes in the model do not necessitate that the estimators be derived separately, and models other than standard statistical models can be estimated with little extra mathematical analysis.

The model for state-level estimates allows covariate information, such as tax and food stamp records, to be treated as response variables themselves. That is, we model the distribution of the covariate, conditional on the proportion to be estimated. The differences in reporting for programs as well as actual differences in the use of programs for people in different places in similarly defined cells are treated as random effects. These, in turn are treated the same way as sampling error in the surveys.

This fully Bayesian model is used to estimate the number of uninsured low-income women within categories defined by age (18-64, 40-64, and 50-64) and race and ethnicity (non-Hispanic Black, non-Hispanic non-Black, and Hispanic) by state. We modeled mutually exclusive age groups (18-39, 40-49, and 50-64) and included the remainder of the population (0-17 and 65 and older) to encompass the entire population. The definition of low-income depends on the state's eligibility criteria for the National Breast and Cervical Cancer Early Detection Program (NBCCEDP). Our estimates include two definitions for low-income, less than or equal to 200 percent and less than or equal to 250 percent of the federal poverty threshold. For completeness, a third income category is used for more than 250 percent of the federal poverty threshold.

Each state has estimates for two sex categories, five age categories, three race and ethnicity categories, and three income categories. This gives 4,950 domains in which to estimate the number of people without health insurance, which are combined into the categories of interest.

The number of people in each income category as well as the proportion with health insurance by ARSH is estimated. The state-level model estimates the number in each IPR category and the proportion insured jointly.

The relationship of the IPR proportions to the ARSH characteristics is modeled with a multiple category logistic regression with normal errors; this is similar to the traditional multinomial regression. The distributions of proportions insured, conditioned on membership in the income categories, are modeled as logistic regressions with normal errors. The data are primarily the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) direct estimates and administrative records data. The CPS ASEC direct estimates are always assumed to be unbiased while the biases of the other data are modeled. The conditional distributions of data are modeled with

normal distributions. All of these data have different functional forms for the modeled variances.

B. Implementation

We use the R software package for data processing. Within R, the R2WinBUGS package runs the WinBUGS software, which performs the MCMC simulation. The R2WinBUGS package is convenient because it facilitates 1) creating data and initial values for WinBUGS, 2) running WinBUGS, and 3) using the simulated data from WinBUGS in R.

WinBUGS software has several positive features for performing the MCMC simulation. Because the algorithm that performs the MCMC simulation is part of the software, the user does not have to spend time writing and debugging the code. Since the user does not have to change the code for each model that is investigated, many different models can be looked at relatively quickly.

WinBUGS has some drawbacks. The degree to which the user can change the MCMC algorithm is limited. Therefore, some models did not perform adequately. For debugging purposes, the error messages that WinBUGS produces are hard to interpret.

Initially, we encountered two significant problems in the course of making our estimates. We initially modeled the error structure of the logistic regressions as multinomial or binomial. When we were unable to get WinBUGS to run this model using non-normal errors, we defaulted to normal errors. The populations are large enough, however, in all but the smallest state/ARSH domains that the normal errors performed well.

The other problem we had was in analyzing the results of the MCMC simulation. WinBUGS produces a text file containing all of the iterations of the Markov chain, for each parameter in the model. To analyze the results, this text file is read into R. However, since our model has about 10,000 parameters, and we generate at least 1,000 iterations for each parameter, the text file is too large for the memory of R. To overcome this problem, we wrote a SAS program that breaks the text file into pieces. Each piece can then be read into R, and analyzed separately.

C. Assessing Results

The fit of the models to the data is very important for the success of these models, and a substantial effort is made in checking the fit. Our primary means of assessing the validity of the models include: examining variances and coefficients of variation (CVs) of the estimates, posterior predictive p-values (ppp-values), and other plausibility checks. The measure of fit was typically ppp-values.[16] In particular, we examine ppp-values

[16] Gelman, A. and Meng, X. (1996), Model Checking and Model Improvement. In Markov Chain Monte Carlo In Practice. Edited by W.R. Gilks, S. Richardson, and D.J. Spiegelhalter; pp. 189-201.

constructed to evaluate the fit with respect to the means and variances, both for global behavior and for behavior within sub-groups.

Examination of ppp-values does not show failures in the model with respect to the expectation of the CPS ASEC or with respect to the variance. Variances and relative variances seem consistent with the conclusion that production of these estimates at this level is feasible. The average posterior CV of the number of uninsured women ages 18-64 with IPR less than or equal to 200 percent of the federal poverty level is about 16 percent. CVs are directly related to confidence intervals (CIs); for a given estimate, smaller CVs correspond with smaller CIs. Smaller CVs and CIs indicate better estimates. The fit of the current model can certainly be improved. In future research, the model will be changed to improve the fit and the predictive power.

VI. County Model

A. Model

Similar estimates are made for the county and state models. However, at the county level, we excluded the categories for race and ethnicity in order to maintain reasonable sample sizes for the domains. The county model estimates insurance coverage by sex, age, and IPR. The sex, age and IPR categories are the same as for the state model.

We estimate the number of uninsured low-income women within age categories (18-64, 40-64, and 50-64) by county. We modeled mutually exclusive age groups (18-39, 40-49, and 50-64) and included the remainder of the population (0-17 and 65 and older) to encompass the entire population. The definition of low-income depends on the state's eligibility criteria for the NBCCEDP. Our estimates include two definitions for low-income, less than or equal to 200 percent and less than or equal to 250 percent of the federal poverty threshold. For completeness, a third income category is used for more than 250 percent of the federal poverty threshold.

Each county has estimates for two sex categories, five age categories, and three income categories. This gives 94,260 domains in which to estimate the number of people without health insurance, which are combined into the categories of interest.

The county model consists of two distinct estimates. The IPR model estimates the number of people in each county/sex/age/IPR domain. The insurance coverage (IC) model is used to estimate the proportion insured within each county/sex/age/IPR category. The estimated proportion of uninsured (from the IC model, by subtracting from one) is multiplied by the estimated number of persons (from the IPR model) to produce the number of uninsured persons for each county/sex/age/IPR.

When the estimation of IPR and health insurance coverage are modeled jointly, as with the state model, the posterior variance of the number insured already incorporates the variance of the number in the IPR category and any covariance between the number in the IPR category and the proportion insured.

The major advantage of separate IPR and insured estimates, as with the county model, is in computational time. The major disadvantage is that it makes estimation of the variance of the estimates more difficult. When the models are separate, we obtain a posterior variance for the number in the IPR category, and a posterior variance for the proportion insured conditional on the number in the IPR category. The estimate of the number insured is the product of the estimate of the number in the IPR category and the estimate of the proportion insured. We must assume independence, or some other particular covariance, to obtain a variance for this product. Our current models assume independence and thus may misrepresent the true variance of the estimate. If this assumption is not true, re-specifying the county-level model may take significant effort.

The basic structure of the county model for number in an IPR category is as follows. We have CPS ASEC estimates of the proportions in the three IPR categories for some of the county/sex/age groups.

For the county IPR model, the following assumptions are made:

- CPS ASEC estimators have normal distributions, whose means are the true proportions in the IPR categories;
- true proportions in the three IPR categories follow a multivariate logistic-normal model;[17]
- variance of the CPS ASEC estimator is inversely proportional to the sample size;[18] and
- variance of the model error is constant.

The IC model is similar. We assume that the proportion insured follows a logistic-normal model with an error term, and that the CPS ASEC estimate is distributed as normal whose mean is the proportion insured.

B. Implementation

We estimate the models using MCMC methods. We have currently implemented the county models as C programs. Compared with WinBUGS, the C program gives us greater control of the MCMC methods used, and allows easier debugging.

We have chosen a version of the MCMC algorithm, which represents a workable compromise between speed and flexibility. Initially, we used the Metropolis algorithm[19] to sample from the full conditional distributions of individual parameters. The Metropolis algorithm has the advantage of being applicable for any density that we can express to within a multiplicative constant. However, convergence can be slow. We therefore analytically derive the full conditional distribution for certain parameters in the model, which results in a large performance increase. This is especially convenient for the vectors of regression parameters, whose joint conditional posterior distributions are multivariate normals that we can sample from relatively easily. These methods have become widespread recently.[20]

While we have sped up the programs considerably, they still run slowly – and thus the process of developing and testing models is slow.

---

[17] This model is a multivariate logistic model except that it includes a normally distributed random effect. In this case, the random effect represents model error.

[18] Fisher, R. and Asher, J. (2000). "Alternate Scaling Parameter Functions in a Hierarchical Bayes Model of U.S. County Poverty Rates", *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: The American Statistical Association.

[19] Metropolis *et al*. (1953). "Equations of State Calculations by Fast Computing Machines", Journal Chem.Phys. 21, 1087-1092.

[20] Gilks, W.R. *et al*. eds (1998). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

C. Assessing Results

We are still assessing the fitness of our models, and investigating alternatives for some of the details. The fit of the models for the IPR and uninsured rate were good overall with some exceptions. There were no indications that the models were misspecified with respect to the logistic regression functions and the distributional assumptions fit well most of the time. The average posterior CV of the number of uninsured women ages 18-64 with IPR less than or equal to 200 percent of the federal poverty level is about 30 percent.

There are areas for further research: the posterior variances are sensitive to the way the variances are modeled. While the variance models seem mostly consistent with the data and with the 'official' variance estimates,[21] we note that the distributional assumptions, including those for normality and for the sampling error variances, are important assumptions and should be examined further. In addition, alternative sets of predictors should be considered. Thus, we should take our posterior distributions as preliminary. The supporting document contains more details.[22]

[21] U.S. Census Bureau (2005). "Source and Accuracy of Estimates for Income, Poverty, and Health Insurance Coverage in the United States: 2004", available at http://www.census.gov/hhes/www/income/p60_229sa.pdf.

[22] Fisher, Robin and Bauder, Mark (2006). "A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html.

PART THREE

DATA

VII. Criteria for Determining Data Adequacy

This evaluation develops and tests the process for estimating the number of eligible women for the National Breast and Cervical Cancer Early Detection Program (NBCCEDP) by state and county. If this information reaches production quality, it will be used by the Centers for Disease Control and Prevention (CDC) in program evaluation and administration. The evaluation uses administrative data, such as the Internal Revenue Service (IRS) tax data, and survey data, such as the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS), to estimate eligibility.

The first step in designing the evaluation is to assess the available data. As described in this part of the report, we reviewed the available data to identify databases that can aid in the estimation of health insurance coverage for low-income women ages 18-64. The basic design for estimating these numbers by county and state was described in Part 2 and in the supporting materials.[23]

A. Overview of Administrative Data

Previous work by the Census Bureau's Small Area Health Insurance Estimates (SAHIE) program has produced county-level small area health insurance estimates. These estimates relied heavily on administrative data such as tax data from the IRS. The method used in the SAHIE program cannot support the requirements of the CDC because of the number of groups that need to be estimated. However, many of the independent variables used or considered for the SAHIE program are the same.

The data are discussed in terms of both their usefulness as a direct estimator and their known problems. The data used in the model and its limitations are discussed. Additionally, data that we chose not to use are listed because in the future some of the data issues may be resolved. This discussion of unused data elements is partially a response to comments received during the review process of the SAHIE program's methodology and estimates. Reviewers asked why certain data sources were not used. In listing them here, we answer those questions.

Administrative databases contain information that is necessary to fulfill the functions of a particular program. The purpose of any administrative database is not research and does not conform to the expected quality of research datasets derived from surveys. Data collection is often done by field staff who make decisions on whether the data field is important in administering the program. In general, fields such as race or ethnicity are not necessary while age is a necessary field. Administrative data are constantly

[23] Fisher, Robin and Bauder, Mark (2006). "A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html.
Fisher, Robin and Riesz, Steven (2006). "A Model for the State-Level Estimation of Insurance Coverage by Demographic Groups", available at http://www.census.gov/hhes/www/sahie/publications.html.

undergoing modifications. For instance, a program administrator may purge data at different times of the year, by state. This older data may be of use to researchers but not of use for the administration of the program.

Population estimates, produced by the Census Bureau, are treated as administrative records for purposes of this report. Population estimates are based on the decennial census and are updated using various administrative records. This information also has problems because it is not designed for our research question.

B. Overview of Survey Data

Data that are based on surveys are constructed for research. Because most surveys are not designed to provide county-level estimates or cover all counties, direct estimates of health insurance for low-income women ages 18-64 are not possible. The decennial census and the American Community Survey (ACS) can provide direct estimates of income categories at the county-level but cannot provide any estimate of health insurance coverage as neither contains questions on this topic.

C. Theoretical Reasons for Data Used

Our review of the types of data needed started with an analysis of the literature pertaining to health insurance coverage at the individual level.[24] When possible, we found data that represented similar information aggregated to the county-level.

This literature review indicated that we need several types of data. First, and most obvious, the literature review indicated that we need the number of people employed in a county, by employer size. Secondly, we need information on income. Third, we need demographic characteristics such as age, race, sex, and Hispanic origin (ARSH) that influence participation in health insurance coverage. Finally, we need information on the different economic conditions of counties that would influence employment, such as urban/rural location, unemployment rates, and information on publicly provided health insurance such as Medicaid.

These types of datum were used, when available, for the SAHIE program's estimates. Since production of the calendar year 2000 estimates, research has been conducted on why Hispanics have the lowest rate of insurance. We have concluded that the number of non-citizen immigrants is an important consideration.[25]

Because this research is not at the unit-level (i.e. person-level), intuitions concerning aggregates may not be valid. This should be kept in mind when intuition is violated. For

[24] Currie & Madrian (1997). Health, Health Insurance and the Labor Market. Handbook of Labor Economics, Volume 3, Edited by O. Ashenfelter and D. Card.
[25] O'Hara, B. (2006). Examining Health Insurance for Hispanics: The Importance of Time and Money. Forthcoming.

instance, size of the employer may not be significant in a model because the distribution of large employers might be similar across counties and states.

D. Determining Data Adequacy

Our assessment of each data element was to:

1. Identify data that might measure theoretical concepts.
2. Check the quality of the data elements in a database.
    a. Review external documentation to identify known problems.
    b. Review internal consistency of variables by checking:
        i. Across domains, including rates.
        ii. Cross tabulations.
3. Check coverage of data element for counties.

E. Overall Conclusions

County-level data that represent many of the important theoretical concepts are available. We were not able to use important data, such as number of employees by firm size, because of suppression of information for some counties.

We describe in Chapter VIII our procedures for acquiring the various databases and processing them to obtain analytical files.

Our assessments of each of the databases considered are the theme of the remaining chapters of Part 3. In Chapter IX, we examine the administrative databases that will be used or were considered: MSIS (Medicaid data), IRS 1040, Food Stamps, Population Estimates, and other extracts that are currently not used in the model. In Chapter X, we examine the survey data that are considered in producing our estimates: CPS ASEC, the decennial census, and ACS. The CPS ASEC is used to construct our dependent variables. The role and constraints of using each of the data are discussed.

VIII. Overview of Data Processing

This section describes the process of acquiring, testing, and manipulating the data that is required for this project.

A. Acquiring External Data

The majority of administrative data that are of interest for making small area health insurance estimates comes from outside the agency. Below is an overview of the process for each administrative database. A new request or a change in scope of the request for data requires lengthier negotiations. The number of staff involved varies for each request depending on: the sensitivity of the data; difficulties in obtaining the data; and the programming/processing effort for both the source agency and the Census Bureau.

1) The Small Area Estimates Branch (SAEB) researches what data are available and how useful they might be through internet searches and inspection of the relevant literature. The Administrative Records Research Staff (ARRS) coordinates with the source agencies and ensures that an agreement will be mutually beneficial.
2) ARRS writes up a draft requirements document describing what data they will acquire from the source agency and any processing ARRS will do once that data are received in-house.
    - ARRS is the only staff that has access to personally identifiable information, such as social security numbers (SSN). ARRS staff remove all personal identifiers (names, addresses, and SSNs) and uses a scrambling algorithm to convert SSNs to a protected identification key.
    - Many of the SAHIE program's data sets are aggregated to the county and state level geographies by ARRS. In particular, the IRS data are completely processed within ARRS and no personal information is passed on to the SAHIE program staff.
3) SAEB reviews draft requirements.
4) ARRS finalizes requirements based on comments received.
5) ARRS writes up an interagency agreement to be signed by representatives of both the Census Bureau and the source agency assuring the source agency that the data will be used only for authorized purposes and confidentiality will be protected.
6) SAEB enters an internal data request in the Administrative Records Tracking System (ARTS). This is the system the Census Bureau uses to keep track of project approval, ensure that data are only used for approved projects, and that everyone working on the project has the appropriate Title 26 training. Title 26 is the statute covering the handling of federal tax information (FTI) and Title 26 awareness training is required for all employees involved with FTI projects.
7) Data are received by ARRS from the source agency.
8) ARRS does some quality checking and processes the data according to the requirements documentation. ARRS delivers the file to SAEB and notes the delivery in ARTS.

9) SAEB receives the data, performs further quality checking and informs ARRS if the data look correct or if there are any problems that need to be addressed.
10) SAEB and ARRS typically meet at least annually to discuss lessons learned and anticipated changes for next year.

B. Acquiring Internal Data

SAEB negotiates with other divisions within the Census Bureau to obtain internal products. A new request for data requires lengthier negotiations. For this project, we have looked into many internal products as detailed below. The order of this process varies somewhat and not every step happens for every request.

1) SAEB researches the data available within the Census Bureau that correspond to concepts related to poverty (income to poverty ratios) or health insurance. SAEB contacts the division that is responsible for the source data. Often, the majority of the negotiation is between branches, not divisions.
2) SAEB meets to discuss any new requirements for these data for this year. SAEB writes up a draft requirements document for review.
3) SAEB extensively reviews the draft requirements document.
4) The source branch reviews requirements documents.
5) SAEB and the source branch meet to discuss any issues and come to agreement on each aspect of the requirements document.
6) SAEB and the source branch baseline the requirements document.
7) The source branch produces data according to the requirements document.
8) The source branch does its own quality checks on the deliverables.
9) SAEB receives the data, performs further quality checking and informs the source branch if the data look correct or if there are problems to be addressed.
10) SAEB and the source branch ideally meet once a year to discuss lessons learned, anticipated changes, etc.

Often, the data needed for new research change during the year. This requires multiple data requests in a year for some datasets.

C. Creating an Analysis File

The modeling team consists of mathematical statisticians who create the formal model and subject matter experts that provide important theoretical inputs for the model. The analysis file consists of these inputs organized into the format used for the statistical model. In this case, the data elements relevant for a county by ARSH and income to poverty ratio (IPR) are included as a line of data (i.e. an observation). Information that is not specific to an ARSH or IPR category is appropriately replicated across observations.

The analysis file is transferred to the mathematical statisticians who program the statistical model, check diagnostics, debug the program after analyzing diagnostics, and

make changes to the model as necessary. This step was described in Part 2 and in the supporting materials.

IX. Assessment of Administrative Data

The SAHIE program uses area-level models that depend on the relevancy, quality, and coverage of administrative data available. The independent variables in an area-level model have to cover the majority of the areas (for example, counties). Administrative data that have national coverage are well suited for this type of model. However, administrative data contain non-sampling errors and do not always cover all areas. In this chapter, we will investigate limitations and errors in the administrative data that are available to us. The following will include a description of the administrative data, which concept they measure for health insurance, their usefulness as a direct measure of health insurance or IPRs, and data issues.

A. Medicaid

Program Description

Medicaid is the largest federal public health insurance program for certain individuals and families with low incomes and low resources. In fiscal year 2005, the average annual enrollment in Medicaid was 44.7 million people with an outlay of $311 billion dollars for medical services.[26] Of those persons with low income and low resources, not all are eligible for medical benefits. Each state administered program sets its own guidelines for eligibility and scope of services. The Medicaid program is typically administered at the local welfare office. Under the Medicaid program, many groups may be covered in a state. These categories typically include welfare participants, those persons ages 65 and over, or persons with a severe disability or blindness. Many state specific programs have been approved by the Centers for Medicare and Medicaid Services (CMS) for covering medical expenses of non-categorical groups. The largest non-categorical programs are for pregnant women, children living in a non-eligible poor family, and persons who are medically needy. All states have CMS approved Medicaid waivers for enrolling women that were screened by the NBCCEDP and determined to have breast or cervical cancer.

The Medicaid program has a cost sharing component with states and the federal government. The cost share is referred to as the Federal Medical Assistance Percentage (FMAP), and is a function of the state's per capita income. Reimbursements of the states' Medicaid expenses are lower for higher income states. The FMAP varies between 50 to 83 percent of Medicaid expenditures.[27] This cost sharing component is important because this influences the extent to which states can offer (or afford) to have extensive Medicaid services particularly during downturns of the economy.

[26] HHS (2005). CMS Financial Report: Fiscal Year 2005.
[27] As discussed earlier, the FMAP is more generous for the NBCCEDP population.

Data issues

The Medicaid data provide a measure of the theoretical concept of public insurance. However, the data cannot be used as a direct estimate of the number of persons insured by public programs or the Medicaid program. The data that we have is called the Medicaid Statistical Information System (MSIS) and is the primary database for Medicaid information. It does not suffice as a direct estimate because:

1. The number of persons listed as eligible for benefits does not match the number of persons that are truly eligible.

Because MSIS is an administrative database, there is not a programmatic reason to purge people from the system who are no longer qualified. Administratively, a person can be in the database but not evaluated for current eligibility until a claim for medical services is filed. At that point, the person will be evaluated for current eligibility. Often, when a person files for benefits the claims representative will flag a duration of eligibility. In this case, a person may be automatically disenrolled after 6 months if the person ceases to file additional claims. At various times it may be beneficial for the administrator to purge persons from the eligible rolls. For instance, an administrator may follow this rule: if a claim has not been submitted for a person in a year, make the person ineligible. Persons can reapply to reestablish eligibility. At a State Health Access Data Assistance Center (SHADAC) conference in 2005,[28] Medicaid administrators said these types of decisions do happen, as well as other factors, and they lead to an overcount of persons with health insurance coverage provided by Medicaid data.

2. Race and Ethnicity are not reliable.

Because race and ethnicity are generally not needed for administrative purposes, the quality of the information varies dramatically across states and counties. Additionally, the information is not entered into the database as race and then ethnicity. Instead, the data field treats Hispanic origin as a race and has a category for multiple races that would include White Hispanic as well as persons considering themselves White and Black. As a result, we are not able to use race and ethnicity from the MSIS data.

3. Not all counties contain reliable data.

There are some counties with suspect counts or missing data. We impute data for these counties. In the model, error in the number of Medicaid participants is incorporated.

4. The distinction between eligible for Medicaid benefits versus eligible for Medicaid and Medicare benefits is not reliable.

The count of persons who are dually eligible for both Medicaid and Medicare is not a reliable data field across states and field offices. As a result, we treat persons with dual eligibility as fully eligible Medicaid participants. We categorize Medicaid participants as

[28] SHADAC conference (2005). Survey and Administrative Data Sources of the Medicaid Undercount.

fully eligible when they have non-restricted benefits or if they are eligible for both Medicaid and Medicare.

5. State differences in Medicaid programs may cause problems in the model.

Further research will be done to determine if separating categorically eligible participants from other participants with full benefits increases the predictive power of the model.

B. Internal Revenue Service 1040 Master File

Program Description

The Internal Revenue Service (IRS) maintains a database of information from the 1040 tax forms. This database is referred to as the Individual Master File (IMF) and is created for each tax year. The IMF contained 130,424,000 tax records and 261,126,000 total exemptions (people of all ages) in 2003.

As discussed in Chapter VIII, before SAEB receives the data individual identifiers are removed from the files and aggregated to the state and county level. Aggregated income and exemptions are available for the SAHIE program.

Data issues

A return (household) is determined to be in an IPR based on reported income and the number and types of exemptions on returns. For instance, a return is poor if the adjusted gross income (AGI) of the return is less than the poverty threshold. These processed data also contain the number of exemptions within IPRs. For the purpose of this study, the tax data contain IPRs of: 0-100, 100-200, and 200 plus. The exemptions are further divided into adult or child exemptions. Calculations are also done to provide a measure of the distribution of the county's taxable income adjusted for family size. The tax data do not provide a method for differentiating between ARSH groups.

The IRS data match important theoretical concepts. First, they provide a measure of the number of persons in an IPR category, by county. This information gives us an independent measure of the sub-groups in the population. Because persons in higher IPR categories are more likely to be insured, this measures the income effect *vis a vis* health insurance. Secondly, the distributional income data of the county as a whole captures the idea that a "rising tide lifts all boats;" counties with a high mean and low variance are likely to have more poor insured than a county with a low mean or high variance.

The IRS data cannot be used as a direct measure of people by IPR. Foremost, the exemptions are primarily by child or adult. Depending on the circumstance, a child exemption is for dependents 0-24. An adult exemption is for all non-child exemptions. In the future, we hope to categorize the IPRs by child, non-elderly adult, and elderly. A tax unit is not the same as a CPS defined family and taxable income is not the same as

CPS family income. These factors result in having the wrong numerator and denominator when calculating poverty rates.

A more fundamental issue with the tax data is nonfilers. CPS survey estimates do not correspond with IRS administrative data as described above. When comparing CPS ASEC data with the IRS data, over 100 percent of the CPS ASEC respondents living in poverty file a tax return while less than 85 percent of higher income respondents (over 200 percent of the poverty threshold) file a return. This result is not intuitive (as income increases, filing rates should increase) and implies a high rate of nonfiling for over two-thirds of our population. We are currently researching this issue and we will model the data to account for the possible data inconsistencies. For these reasons, we also expect the distributional variables to not represent the true distribution of income.

As the discussion of nonfilers implies, an IPR from CPS is not the same as an IPR measurement from tax data. CPS income is based on adding different types of income and CPS poverty thresholds are based on the number of persons in a related family. For tax data, the income measure is AGI and the poverty threshold is based on the number of exemptions in a tax unit. These differences should lead to more people being tax-poor than CPS-poor.

We will use the data as described and adjust for known biases.

C. Food Stamps

Program Description

The U.S. Food Stamp Program (FSP) is the nation's largest nutrition program for low-income Americans and a source of demand for the products of American farmers and food industries. The program provides benefits with electronic debit cards, which participants may use to buy food from eligible retailers. The program served about 21.3 million low-income Americans on average each month in fiscal year 2003, with a United States Department of Agriculture (USDA) outlay of about $23.9 billion.[29] The FSP is typically administered at the local welfare office. The federal government pays for all of the benefits and part of the administrative costs.

The Food Stamp Program is the one low-income assistance program that is uniform in its eligibility requirements and benefit levels across states, except for Alaska and Hawaii, where benefit levels and income eligiblility requirements are higher. While the definitions of income and household composition are different from those used in the official measure of poverty, a household's eligibility for the program is determined by a standard that is tied to the poverty level. With the exception of Alaska and Hawaii, the household's gross income needs to be below 130 percent of the official poverty threshold. State-to-state variation occurs due to differing ways that the program is implemented. For instance, the effort for a family to continue to receive food stamps benefits depends

[29] http://www.ers.usda.gov/Briefing/FoodStamps

on the state specific rules on recertification and states can decide on what resources and income to count in determining income.

Data Issues

We obtain counts of the number of people participating in the food stamp program from the USDA, Food and Nutrition Service (FNS).[30] For counties we use counts of participants for the month of July in the estimation process. In a few cases, however, the states were able to provide data only for other reference periods. For states, we use a 12 month average from July of the reference year to June of the following year.

The food stamp data correspond to the idea of counting the number of people in poverty or near poverty. These data cannot be used as a direct estimate of poverty. First, the IPR category for eligibility is not the same as poverty (less than 130 percent as opposed to less than 100 percent). Second, most persons applying for the program have to have low resources and pass a second income screening process. Third, many persons categorically qualify for food stamps because of their eligibility for other programs (Temporary Assistance for Needy Families and Supplemental Security Income) and are exempt from all income and asset tests. Fourth, many people never apply that are qualified for benefits. The percent of non-applicants varies by state from 39 percent to 81 percent.[31] Finally, the data are not available by ARSH groups.

We will use the data as described and adjust for known biases.

D. Population Estimates

Program Description

The Census Bureau's Population Estimates Program (PEP) produces estimates of the county populations, starting with the base populations from Census 2000. With each new issue of July 1 estimates, estimates are recalculated for years back to the last decennial census. Previously published estimates are considered out-of-date and archived. Estimates from PEP are the basis for the population weights in Census Bureau surveys.

The Census Bureau develops county population estimates with a demographic procedure called an "administrative records component of population change" method. A major assumption underlying this approach is that the components of population change are closely approximated by administrative data in a demographic change model. In order to apply the model, Census Bureau demographers estimate each component of population change separately. For the population residing in households the components of

[30] We make the state and county food stamp data, used in our programs, available at http://www.census.gov/hhes/www/saipe/tables.html
[31] http://www.mathematica-mpr.com/publications/pdfs/fns02rates.pdf

population change are births, deaths, and net migration, including net international migration.[32]

Data issues

In our small area estimation method, we require estimates of the ARSH population that cover all counties to make a final estimate. The CPS data are controlled to population estimates at the state-level but not by the same ARSH categories so inconsistencies will arise. Further inconsistencies will arise at the county-level because CPS is not controlled to this level of geography. In other words, county totals for an ARSH group will generally be different when comparing CPS and population totals.

To calculate the final estimates, population numbers are used to convert the modeled ratio results to numbers. The population estimates are also a component in the calculation of variance.

E. Other Administrative Data, Rejected

1. State Children's Health Insurance Program

Program description

CMS provides health insurance coverage for low-income children without health insurance through the State Children's Health Insurance Program (SCHIP). In 2004, SCHIP had 6.2 million enrollees and had 5.4 billion in expenditures for health services.[33] SCHIP is similar, but separate, from Medicaid. The program prioritizes children in families that earn too much to receive Medicaid benefits but do not have sufficient resources to purchase health insurance. Congress determined the minimum standard for lacking resources. If the family income is at or below 200 percent of the federal poverty threshold or the family income is less than 150 percent of the state's Medicaid eligibility threshold then the child is eligible for SCHIP benefits.

The funding for SCHIP is complicated. The federal government allocates state funding according to a statutory formula that is based on the "Number of Children" and the "State Cost Factor." The Number of Children factor is 50 percent of the low-income uninsured children. The State Cost Factor is the annual wages in the health care industry in the state. Each state's allotment represents the total possible expenditures that the state can receive for qualifying matching funds. States receive an enhanced federal matching rate that is equal to 70 percent of their Federal Medical Assistance Percentage (FMAP) plus 30 percentage points for covered medical expenses. The total matching rate cannot exceed 85 percent; this has the affect of substantially raising the federal contribution for states with higher incomes.

[32] Complete documentation of these methods is available at http://www.census.gov/popest/counties.
[33] CMS (2005). FY 2004 SCHIP Annual Enrollment Reports.

The state can either use SCHIP funds to (Option 1) expand Medicaid eligibility to children who previously did not qualify for the program; (Option 2) design a separate SCHIP entirely separate from Medicaid; or, (Option 3) combine both the Medicaid and separate program options.

Data issues

SCHIP measures the theoretical concept of public insurance for non-poor low-income children. It could also serve as an estimate for non-poor low-income families. However, it cannot be used as a direct estimate. The data that we currently have are from the MSIS and are the primary database for Medicaid information. For states that choose to fund their SCHIP program exclusively through the expanded Medicaid option, MSIS captures all of the SCHIP information. For states that design a separate program, MSIS captures none of the SCHIP children. For combination programs, MSIS captures only children in the expanded Medicaid option of SCHIP.

It does not suffice as a direct estimate because (1) of the same limitations as Medicaid listed above and (2) we do not have any or all of the information for states choosing options 2 or 3. Therefore, all SCHIP information contained within MSIS is ignored. However, we are looking into obtaining data containing the remainder of SCHIP participants (the SCHIP Enrollment Data System).

2. County Business Patterns

Description of the data

The County Business Patterns (CBP) data are extracted from the Business Register, the Census Bureau's file of all known single and multi-establishment companies. The Annual Company Organization Survey and quinquennial Economic Censuses provide individual establishment data for multi-location firms. Data for single-location firms are obtained from various programs conducted by the Census Bureau, such as the Economic Censuses, the Annual Survey of Manufactures, and Current Business Surveys, as well as from administrative records of the IRS, the Social Security Administration, and the Bureau of Labor Statistics. Data are excluded for self-employed persons, employees of private households, railroad employees, agricultural production workers, and for most government employees.

Data issues

At the individual level, there is a clear relationship between firm size and the offer of health insurance. At the county-level, this concept would correspond with the number of people employed at private establishments of different size. These data would serve that purpose if important categories were not suppressed. Ideally, we want the data a bit differently than are publicly available: 1) the number of persons working in the different size firms, not the number of different size firms; 2) a separation of service and all other industries. We will obtain unsuppressed data with the above specification, but will not

receive them in time for this deliverable. We hope to incorporate these data in the next version of our models. At this point, the level of suppression of the public use files makes these data not usable.

Based on the research conducted on suppressed data, these data have not been a good predictor of health insurance. To be a good predictor in a county-level model, there has to be significant variation in the ratio of employees in different size establishments across counties. We hope that the data will be useful when we have unsuppressed counts of employees by firm size.

3. Quarterly Census of Employment and Wages Program

Description of the data

The Bureau of Labor Statistics gathers data from the State Employment Security Agencies to produce the Quarterly Census of Employment and Wages (QCEW) data file. The QCEW program produces employment and wage information for workers, by the North American Industry Classification System (NAICS), covered by Unemployment Insurance or Unemployment Compensation. The database includes data on monthly employment by ownership sector aggregated to the county-level.

Data issues

Like the CBP data, much of the publicly available data are suppressed to protect the privacy of the reporting businesses. The public data were used to analyze the effect of federal and state employment. The amount of suppression made it difficult to assess the use of these data but they have advantages over the CBP because they contain private sector as well as public sector employment. We are trying to obtain unsuppressed data for evaluation purposes. We expect this to be more of a long-term project because the data are from another agency.

4. Federal / State Government Employment and Payroll

Description of the data

Data in these files are based on information obtained in the Annual Survey of Government Employment and Payroll. Federal government data were compiled by Census Bureau staff from records of the U.S. Office of Personnel Management. Approximately one-half of the state governments provided data from central payroll records for all or most of their agencies and institutions. Data for agencies and institutions for the remaining state governments were obtained by mail canvass questionnaires. Local government data were generally requested by mail canvass questionnaires except for the following: elementary and secondary school system data in Florida, North Carolina, North Dakota, and Washington were supplied by special arrangements with the state education agency in each of these states.

Data issues

Federal, state and employees in the local schools are likely to be insured due to unionization. This information would likely be a good predictor of health insurance coverage. However, available databases do not have county-level information and this information is by the place of employment not the county of residence. We are currently negotiating for this information. The QCEW data are probably a superior source of government employees, but they are less accessible than Government Employment and Payroll data.

5. Regional Economic Information System

Description of the data

The Bureau of Economic Analysis produces a collection of data in their Regional Economic Information System (REIS). This system integrates information from over 300 surveys and administrative databases to produce county, state, regional, and national economic statistics. Of the data collected, unemployment rates, average wages, and the percent of payroll spent on benefits are of particular interest. Unemployment rates reflect varying levels of potential employer-provided health insurance. Payroll figures reflect overall generosity of employers. Average low wages should indicate a lower level of insured rates because employees will have a harder time purchasing insurance and employers may be less likely to offer insurance.

Data issues

Like other public data, the REIS data have been subjected to discloser rules and some of the lower level data have been suppressed. However, even when using the suppressed data, it appears generosity of benefits varies across counties and provides a good predictor of county-level insurance coverage. However, the level of suppression is relatively high. We expect obtaining unsuppressed data to be more of a long-term project because the data are from another bureau.

X. Assessment of Survey Data

The SAHIE program uses survey data for many purposes. The Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) provides the rate of health insurance coverage and IPRs for counties that contain sampled households. The American Community Survey (ACS) and the decennial census provide information on: IPRs; the counties general ability to afford health insurance; and the number of non-citizen immigrants, who tend to have much higher uninsured rates. At this time, the county model uses the decennial census to predict the number of persons in an ARSH/IPR category instead of the ACS.

A. Annual Social and Economic Supplement of the Current Population Survey

Description

The Current Population Survey (CPS) is a monthly survey of about 60,000 eligible households conducted by the Census Bureau for the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is selected to represent the civilian, noninstitutional population.

The Annual Social and Economic Supplement (ASEC) contains the basic monthly demographic and labor force data described above, plus additional data on work experience, income, noncash benefits such as health insurance, and migration. The sample size for the CPS ASEC was increased to approximately 85,000 eligible households in 2002. The primary goal of the sample expansion was to produce more reliable estimates of low-income children without health insurance for the State Children's Health Insurance Program (SCHIP) through reduced variances. Although the SCHIP sample expansion was specifically designed to produce better estimates of children's health insurance at the state-level, other state and national estimates are also improved.

The health insurance data combined with ARSH characteristics are based on the CPS ASEC. Health insurance includes: private health insurance, Medicare, Medicaid, SCHIP, Military Health Care, and state specific plans for low-income individuals. Using the CPS definition of health insurance, Indian Health Service itself is not included as health insurance. This treatment of Indian Health Service is consistent with Government Accountability Office (GAO) findings[34] and the Native American Breast and Cervical Cancer Treatment Technical Amendment Act of 2001.

CPS ASEC data are used to determine the official poverty rate. All income questions refer to the previous calendar year and family size is determined at the time of the interview. The IPR is created by dividing family income by the appropriate poverty threshold.

[34] GAO (2005). Indian Health Services: Health Care Services Are Not Always Available to Native Americans.

The data are aggregated to a county-level and adjusted to make each county self-representing. Within each county, the population is categorized by ARSH and IPR.

Data issues

With the SCHIP expansion, it is possible to make state-level estimates of uninsured low-income women using a multi-year average. The sample size is too small for reliable direct estimates of the uninsured, by ARSH, at the state and county-levels. At the county-level, about two-thirds of counties do not contain any sample. Only 29 percent of all counties have any uninsured low-income women. The sample size issue becomes more problematic as the sample is categorized by ARSH characteristics. Sparse data lead to high variances. By using small area estimate techniques, the state estimates will be improved and reliable county estimates become possible.

B. Census 2000

Description

A limited number of questions were asked of every person and at every housing unit in the United States. An expanded set of questions was asked of a sample of the total population (about 1 in 6). The Census 2000 does not contain questions on health insurance coverage. Estimates of the characteristics of the population and housing of communities throughout the United States are produced. These characteristics include demographic, social, and economic information about individuals and households; also, they include physical and financial characteristics of housing units. Of specific interest to this project, the Census 2000 contains information on ARSH, family income, and family size. This information is used to establish IPRs by ARSH. In the Small Area Income and Poverty Estimates (SAIPE) program, the decennial census information has been a valuable predictor of poverty in non-decennial years. The distribution of education and the number of non-citizen immigrants by ARSH is also a potential predictor of the number of persons with health insurance coverage.

Data issues

Census 2000 data provide accurate information on IPRs. However, they cannot provide a direct estimate of IPRs because 1) the measures of income (and therefore poverty) are different than in the CPS ASEC[35] and 2) Census 2000 information refer to income year 1999 and is less relevant as the economy changes during the intercensal years.

We will explore the use of decennial census information in our county and state model to predict the number of persons in IPR/ARSH groups. In future research, these data and information on non-citizens and education will be included in the insurance coverage

[35] Differences between the CPS ASEC and Census 2000 measures of poverty result from the use of different questions to collect income information, data collection techniques, and sampling design.

portion of the model as indicators of the overall ability to pay for health insurance coverage.

C. American Community Survey

Description

The American Community Survey (ACS) is a large, continuous demographic survey that will replace the decennial census long form. The ACS does not currently contain questions on health insurance coverage. The ACS produces annual and multi-year estimates of the characteristics of the population and housing of communities throughout the United States. These characteristics include demographic, social, and economic information about individuals and households; also, they include physical and financial characteristics of housing units.

The ACS does not provide official population counts. That information is collected once every ten years by the decennial census short form; the counts are updated every year by the Census Bureau's intercensal Population Estimates Program (PEP).

The ACS measures income and poverty on a rolling schedule. Each month households are asked questions concerning family income over the previous 12 months. IPRs are determined with this information. Full implementation occurred in 2005. ACS does not contain information on health insurance.

Data issues

ACS can produce a direct estimate for IPRs by ARSH when the data become available. However, the design of ACS does not support an estimate based on one year of data for small counties. Counties with between 20,000 and 65,000 total population will be reported as 3-year averages and counties with less than 20,000 total population will be reported as 5-year averages. Income and poverty are measured differently between the CPS ASEC and ACS. The surveys have different recall periods, different questions are used to collect the income data, and the data are collected using different methods (phone interview vs. mail out/mail back). Because the surveys produce different answers, for like concepts, it is not clear how these data can be incorporated into this project. At this point, the ACS data will not be used but investigation of this data source will continue when data are available from the full implementation of the survey. When there is sufficient sample for all counties, the ACS will supplant the information gathered from the decennial census.

PART FOUR

U.S. CENSUS BUREAU STANDARDS

XI. Discussion of Race

This chapter discusses the assignment of race and ethnicity categories for this evaluation and changes we plan to make for the future. When tabulating population and the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) estimates by race and ethnicity we need to assign a race and ethnicity for all respondents based on their response.

In 1977 the Office of Management and Budget (OMB) issued a directive establishing standards for four minimum race categories: American Indian or Alaskan Native, Asian or Pacific Islander, Black, and White and two ethnicity categories: Hispanic origin and Not of Hispanic origin.

In 1997 the OMB revised the standards for data on race and ethnicity creating the following minimum race categories:

- American Indian or Alaska Native;
- Asian;
- Black or African American;
- Native Hawaiian or Other Pacific Islander; and
- White.

OMB now allows respondents who self-identify to select more than one race creating the category two or more races. The two categories for ethnicity are: Hispanic or Latino and Not Hispanic or Latino. Hispanics and Latinos may be of any race.

During early dicussions the Centers for Disease Control and Prevention (CDC) requested that we follow the method they were currently using to assign race and ethnicity. That method keeps the categories mutually exclusive and eliminates the two or more races category by "promoting-up" respondents in this category based on minority status. For example, a person who listed Black or African American and White was assigned to the Black or African American category. A person who listed American Indian or Alaska Native, Black or African American, and White was assigned to the American Indian or Alaska Native category. Further details will not be given as this method will not be used in the future. Switching from the the "promoting-up" method should not have a significant impact on our results.[36]

Since we requested that the data be tabulated according to the "promoting-up" method this assignment of race and ethnicity is used in the feasibility study. For research purposes, we further collapsed the data into the following categories: non-Hispanic Black, non-Hispanic non-Black, and Hispanic.

When the Census Bureau met with the CDC in October 2005 the issue of race and ethnicity was addressed. The CDC no longer prefers the "promoting-up" method and

[36] Mills, Robert and Bhandari, Shailesh (2003). Health Insurance Coverage in the United States: 2002. U.S. Census Bureau, Current Population Reports, P60-223. Washington, DC: U.S. Government Printing Office.

would like to keep the category two or more races. This is in agreement with Census Bureau standards. During internal discussions among Census Bureau staff, the Small Area Health Insurance Estimates (SAHIE) program was strongly recommended not to use the "promoting-up" method.

Several options were discussed. We debated using categories of "race alone" or using "race alone or in any combination." Census Bureau reports such as the P-60 series "Income, Poverty, and Health Insurance Coverage in the United States" are moving in the direction of reporting race alone so this option was selected. We were strongly advised to collapse other races that were not Hispanic because of the small sample sizes across most states. The following race and ethnicity categories will be used for our next data request:

- White alone, not Hispanic;
- Black or African American alone, not Hispanic;
- Other, not Hispanic (American Indian or Alaska Native alone, not Hispanic; Asian alone, not Hispanic; Native Hawaiian or Other Pacific Islander alone, not Hispanic; and Two or More Races, not Hispanic); and
- Hispanic.

The above categories will be requested for the population estimates and the CPS ASEC estimates beginning with the 2003 CPS. This is the first year where CPS respondents could identify themselves in more than one race group. Prior to 2003, CPS respondents selected a single race group. When averaging 3 years of the CPS ASEC we will be combining single race data with race alone data until all 3 years have race alone data. This should not have a significant impact. In Census 2000, 2.4 percent of the population selected two or more races. Of the persons who reported multiple races, the most common combination was White and one or more races.[37]

[37] Grieco, Elizabeth and Cassidy, Rachel (2001). Overview of Race and Hispanic Origin 2000. Census 2000 Brief, C2KBR/01-1. Washington, DC: U.S. Government Printing Office.

XII. Publishing Estimates from the NBCCEDP Project

The Census Bureau has a long history of high standards in its data products. These standards hold true for both special tabulations done for specific agencies or groups and data published directly on the Census Bureau's web site. These standards are a reflection of the Census Bureau's commitment to providing quality data products and the Office of Management and Budget's (OMBs) **"**Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies."[38]

The release of the first ever set of estimates under the SAHIE program, which was partially supported by the CDC, included substantial internal and external review. Early in the development process, methodology was vetted through the Census Advisory Committee of Professional Associations. Later, after the methods were enhanced, based partially on advice from the Advisory Committee, the estimates and methods were reviewed internally by Census Bureau senior technical staff and upper management. External experts in the field of health insurance coverage also reviewed the methods and estimates. Based on a recommendation from the OMB, the Census Bureau also vetted the estimates and methods to other federal agencies. The University of Minnesota's State Health Access Data Assistance Center (SHADAC) conducted a review separate from government agencies.[39] SHADAC provides technical advice to states that use health insurance data or advice for states to create their own surveys concerning health insurance.

What does this mean relative to the National Breast and Cervical Cancer Early Detection Program (NBCCEDP)? The CDC has indicated that they would like any final estimates the Census Bureau can produce under this project to be published on the Census Bureau's web site. To meet this requirement, the Census Bureau will subject any estimates to comprehensive quality review. Because we have already undergone substantial internal and external review with the base SAHIE estimates, the level of external review for estimates relating to the NBCCEDP would not require as extensive a review. However, the publication of estimates would be required to meet other Census Bureau quality standards. In particular, we note that there is a certain level of information that must be published with any estimates. This documentation of methods and assessment of the precision and un-biasness of the estimates must be a part of any release and requires more effort than may be required to deliver data to the CDC for publishing on the CDC site.

Although a web site exists for the SAHIE estimates, publishing the NBCCEDP project estimates would require further development of the web site that would minimally include documentation of the estimates and an additional data base to access and display the data. Developing, testing, and maintaining this part of the web site would be included in the cost of publishing the estimates.

---

[38] Information on the specifics of the Census Bureau quality guidelines and the OMB directive are available on the Census Bureau web site at: http://www.census.gov/quality.

[39] The results of that review and the Census Bureau's response are available on our web site at: http://www.census.gov/hhes/www/sahie/index.html.

Appendix A: Acronyms

Abbreviation	Definition
ACS	American Community Survey
AGI	Adjusted Gross Income
ARRS	Administrative Records Research Staff
ARSH	Age, Race, Sex, and Ethnicity
ARTS	Administrative Records Tracking System
ASEC	Annual Social and Economic Supplement
CBP	County Business Patterns
CDC	Centers for Disease Control and Prevention
CMS	Centers for Medicare and Medicaid Services
CPS	Current Population Survey
CV	Coefficient of Variation
DAG	Directed Acyclic Graph
FMAP	Federal Medical Assistance Percentage
FNS	Food and Nutrition Service
FSCPE	Federal and State Cooperative Program for Population Estimates
FSP	Food Stamp Program
FTI	Federal Tax Information
GAO	Government Accountability Office
GLIM	Generalized Linear Models
IC	Insurance Coverage
IMF	Individual Master File
IPR	Income to Poverty Ratio
IRS	Internal Revenue Service
MCMC	Markov Chain Monte Carlo
MSIS	Medicaid Statistical Information System
NAICS	North American Industry Classification System
NBCCEDP	National Breast and Cervical Cancer Early Detection Program
OMB	Office of Management and Budget
PEP	Population Estimates Program
PPP-Values	Posterior Predictive P-Values
REIS	Regional Economic Information System
SAE	Small Area Estimates
SAEB	Small Area Estimates Branch
SAHIE	Small Area Health Insurance Estimates
SAIPE	Small Area Income and Poverty Estimates
SCHIP	State Children's Health Insurance Program
SDC	State Data Center
SHADAC	State Health Access Data Assistance Center
QCEW	Quarterly Census of Employment and Wages
USDA	United States Department of Agriculture