University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2006 *Paper* 58

Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening

Trivellore E. Raghuanthan[*] Dawei Xie[†] Nathaniel Schenker[‡]

Van Parsons[**] William W. Davis[††]

Kevin W. Dodd[‡‡] Eric J. Feuer[§]

[*] University of Michigan, teraghu@umich.edu
[†] University of Pennsylvania
[‡] National Cancer Institute
[**] National Center for Health Statistics
[††] National Cancer Institute
[‡‡] National Cancer Institute
[§] National Cancer Institute

Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening

Trivellore E. Raghuanthan, Dawei Xie, Nathaniel Schenker, Van Parsons,
William W. Davis, Kevin W. Dodd, and Eric J. Feuer

Abstract

Cancer surveillance requires estimates of the prevalence of cancer risk factors and screening for small areas such as counties. Two popular data sources are the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey conducted by state agencies, and the National Health Interview Survey (NHIS), an area probability sample survey conducted through face-to-face interviews. Both data sources have advantages and disadvantages. The BRFSS is a larger survey, and almost every county is included in the survey; but it has lower response rates as is typical with telephone surveys, and it does not include subjects who live in households with no telephones. On the other hand, the NHIS is a smaller survey, with the majority of counties not included; but it includes both telephone and non-telephone households and has higher response rates. A preliminary analysis shows that the distributions of cancer screening and risk factors are different for telephone and non-telephone households. Thus, information from the two surveys may be combined to address both nonresponse and noncoverage errors. A hierarchical Bayesian approach that combines information from both surveys is used to construct county-level estimates. The proposed model incorporates potential noncoverage and nonresponse biases in the BRFSS as well as complex sample design features of both surveys. A Markov Chain Monte Carlo method is used to simulate draws from the joint posterior distribution of unknown quantities in the model based on the design-based direct estimates and county-level covariates. Yearly prevalence estimates at the county level for 49 states, as well as for the entire state of Alaska and the District of Columbia, are developed for six outcomes using BRFSS and NHIS data from the years 1997-2000. The outcomes include smoking and use of common cancer screening procedures. The NHIS/BRFSS combined county-level estimates are substantially different from those based on BRFSS alone.

Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening

Trivellore E. Raghunathan, Dawei Xie,
Nathaniel Schenker, Van Parsons,
William W. Davis, Kevin W. Dodd, and Eric J. Feuer [1]

Abstract

Cancer surveillance research requires estimates of the prevalence of cancer risk factors and screening for small areas such as counties. Two popular data sources are the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey conducted by state agencies, and the National Health Interview Survey (NHIS), an area probability sample survey conducted through face-to-face interviews. Both data sources have advantages and disadvantages. The BRFSS is a larger survey, and almost every county is included in the survey; but it has lower response rates as is typical with with telephone surveys, and it does not include subjects who live in households with no telephones. On the other hand, the NHIS is a smaller survey, with the majority of counties not included; but it includes both telephone and non-telephone households and has higher response rates. A preliminary analysis shows that the distributions of cancer screening and risk factors are different for telephone and non-telephone households. Thus, information from the two surveys may be combined to address both nonresponse and noncoverage errors. A hierarchical Bayesian approach that combines information from both surveys is used to construct county-level estimates. The proposed model incorporates potential noncoverage and nonresponse biases in the BRFSS as well as complex sample design features of both surveys. A Markov Chain Monte Carlo method is used to simulate draws from the joint posterior distribution of unknown quantities in the model based on

[1] Trivellore E. Raghunathan is a Professor of Biostatistics and a Research Professor at the University of Michigan, Dawei Xie is an Instructor in the Department of Biostatistics and Epidemiology at the University Pennsylvania, Nathaniel Schenker is the Senior Scientist for Research and Methodology at the National Center for Health Statistics (NCHS), Van Parsons is a Mathematical Statistician at NCHS, William W. Davis is a Mathematical Statistician at the National Cancer Institute (NCI), Eric J. Feuer is the Chief of the Statistical Research and Application Branch at NCI, and Kevin W. Dodd is a Mathematical Statistician at NCI. This research was supported through contracts 263-MQ-320576 and 263-MQ-511439 from NCI and grants DMS-9803720 and SES-010694 from the National Science Foundation to the University of Michigan. The authors are listed by institution, and the first author listed under each institution is the co-lead investigator. The views expressed in this paper are those of the authors and do not necessarily represent the views of NCHS or NCI.

the design-based direct estimates and county-level covariates. Yearly prevalence estimates at the county level for 49 states, as well as for the entire state of Alaska and the District of Columbia, are developed for six outcomes using BRFSS and NHIS data from the years 1997-2000. The outcomes include smoking and use of common cancer screening procedures. The NHIS/BRFSS combined county-level estimates are substantially different from those based on BRFSS alone.

Key Words: Hierarchical model, Gibbs sampling, Complex sample survey, BRFSS, NHIS, Simulation, Cancer screening, Mammography, Smoking, Pap smear

1 Introduction

1.1 The Need for Small-Area Estimates

Cancer surveillance research requires estimates of the prevalence of various characteristics for small areas. Often the small areas are counties or collections of counties defined as Health Service Areas. The characteristics of interest include life-style variables (e.g., smoking, dietary habits, physical activity, and obesity), economic status (e.g., education and income), and health care utilization (e.g., insurance use and cancer screening practices). Small-area estimates are used by researchers in trend analysis, in predicting future cases, in risk analysis, and in investigating relationships between risk factors and cancer outcomes such as incidence, mortality, and survival. Improved small-area estimates may yield improved predictions and risk estimates.

Trends in the characteristics of interest have policy implications at both the national and the small-area levels. For example, differential rates of cancer screening use in the United States (U.S.) by age, race, health status, and socioeconomic factors have been well documented (Breen et al. 2001, Swan et al. 2003). This recognition has stimulated a greater focus on intervention research that targets populations with low utilization rates. Although successful strategies to increase cancer screening among underutilizing populations have recently been reported, there has been unevenness in the targeting of intervention research, resulting in gaps in coverage. These gaps are evident geographically, which may warrant further investigation of the need for tailoring intervention research as well (Legler et al. 2002).

Recently, county-level risk factor prevalence rates have been used to predict the number of new cancer cases in the next year by state, gender, cancer type, and race/ethnicity (Pickle et al.

2001). The approach utilized results from the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Result (SEER) program of tumor registry data, which currently covers approximately 26% of the U.S population. An ecological regression equation was developed using incidence data from SEER and risk factor estimates to predict incidence in non-SEER areas at the county level. Thus, obtaining accurate and precise small-area estimates for cancer risk factors is an important problem.

1.2 The Behavioral Risk Factor Surveillance System and the National Health Interview Survey

A popular data source for obtaining small-area estimates has been the Behavioral Risk Factor Surveillance System (BRFSS), an ongoing telephone survey of the health behaviors of U.S. adults, established in 1984 by the Centers for Disease Control and Prevention (CDC). The BRFSS was designed to provide state-specific estimates of the prevalence of risk behaviors. Its strength for small-area estimation is the large sample taken in each state. In 1997, the state sample sizes ranged from 1,505 people in the District of Columbia (DC) to 4,923 people in Idaho (Iachen et al. 1999). Most of the counties in the U.S. are included in the BRFSS sample as well. The total sample size for the U.S. increased each year from 1997 (over 130,000 people) through 2000 (over 180,000 people). (National sample sizes are available online at www.cdc.gov/brfss/technical_infodata/.) With the BRFSS, the states have the freedom to implement their own sampling protocols, though some features have been standardized. In 1997, 29 states (including DC) used variations of the Waksberg-type multistage cluster design (Waksberg 1978), while 22 states used some type of list-assisted design (e.g., Lepkowski 1988).

Coverage can be a problem for telephone surveys such as the BRFSS. The degree of bias in surveys that exclude households without telephones is a function of the percentage of households without telephones, the magnitude of the difference between owners and non-owners of telephones on the particular outcome, and the adjustment technique used, if any. Based on the 1990 decennial census, only 5.2% of the occupied households did not have telephones; however, 12.6% of the occupied homes in Mississippi did not have telephones. Based on the 2000 census, the number of households without telephones at the county level ranged from 0.4% and 46.1%. Coverage also varies substantially among racial/ethnic groups, and coverage is substantially below the national average in households with the lowest per capita income. In 1994, for example, an estimate, based on the National Health Interview Survey (NHIS) of CDC's National Center for Health Statistics (NCHS), of non-telephone coverage for blacks below the poverty level was 21.3% (Anderson et al. 1998). Thus, for outcomes where there is a substantial difference between those below and above the poverty line, telephone survey estimates could be seriously biased.

Another source of bias for list-assisted designs arises due to noncoverage of unlisted numbers. Furthermore, for efficiency reasons, sometimes blocks of numbers that have fewer than a pre-specified number of residential lines are excluded from the sampling frame. These biases can be substantial, especially if the prevalence of risk factors is correlated with having a telephone. Thus, the BRFSS alone may not be ideal for developing small-area estimates.

An alternative data source is the NHIS, a nationally representative, stratified, multistage, area probability sample of households that collects information based on face-to-face interviews. A new sample design is implemented following each decennial census. The 1995-2004 NHIS was designed to produce estimates for the nation, for each of four census regions, and within regions by areas determined by metropolitan status. The total number of households sampled each year in the NHIS is approximately 40,000. For the outcomes considered in this paper, data are available for a sample of adults from these households (the NHIS "adult sample"), the number of which is roughly the same as the number of households. Although the survey samples from all of the states and DC each year, it is not designed to produce reliable direct state-level estimates for every state (Botman et al. 2000). Moreover, only about 25% of the counties in the U.S. are included in the sample, so small-area estimates obtained solely from the NHIS may be unreliable.

The advantages of the NHIS are that it includes both telephone and non-telephone households, and that it has higher response rates than does the BRFSS. For example, the response rates for the adult sample in the NHIS, which account for nonresponse (both refusals and other types of nonresponse) by families within sampled households as well as nonresponse by adults sampled from responding families, were 80.4%, 73.9%, 69.6%, and 72.1% in 1997 through 2000, respectively (National Center for Health Statistics 2000a,b, 2002a,b). In contrast, the BRFSS state-level (including DC and Puerto Rico) response rates (calculated via the method suggested by the Council of American Survey Research Organizations) had median values of 62.1% in 1997, 59.1% in 1998, 55.2% in 1999, and 48.9% in 2000 (Centers for Disease Control and Prevention, 2001). The state-level response rates ranged from 41.3% (Hawaii) to 88.9% (Puerto Rico) in 1997, 32.5% (Deleware) to

76.7% (Puerto Rico) in 1998, 36.2% (Texas) to 80.8% (Minnesota) in 1999, and 28.8% (New Jersey) to 71.8% (Montana) in 2000.

1.3 A Project to Combine Information from the BRFSS and the NHIS

Fortunately, several questions are common between the BRFSS and the NHIS. Furthermore, the NHIS also asks, “Is there at least one telephone INSIDE your home that is currently working?” Thus, one strategy is to combine information from both surveys to obtain small-area estimates correcting for both noncoverage due to not having telephones and nonresponse. There are two possible approaches for combining information. The first approach, discussed in Elliott and Davis (2005), uses only publicly available NHIS data, which contain no geographic identifiers beyond the four U.S. census regions of residence (Northeast, Midwest, South, and West) and seven urban-rural categories (ranging from metropolitan statistical area greater than 5,000,000 to non-metropolitan statistical area), to calibrate the BRFSS estimates to adjust for nonresponse and noncoverage. Since much gain in efficiency might result in using the actual county identifiers from the NHIS, a collaborative project between NCI, NCHS (and its parent agency, CDC), and the University of Michigan, was undertaken to develop small-area estimates based on combining information from both surveys using data that include county identifiers. For successful completion, this project required extensive collaboration and cooperation among the institutions involved.

Prevalence rates for six outcomes, including four cancer risk factors and the use of two types of cancer screening, were of primary interest for the first phase of this project, and they are listed in Table 1. The four cancer risk factors are gender-specific smoking, current and ever, among those who are at least 18 years of age; and the two types of cancer screening are mammography during the past two years among women who are at least 40 years of age and pap smear testing during the past three years among women who are at least 18 years of age. Estimates were obtained for these 6 outcomes for 3,114 small areas: 3,112 counties in 49 states, and the entire state of Alaska and DC. The data from both surveys for the years 1997-2000 were used to obtain annual estimates. For this project, 20 county-level covariates, which are also listed in Table 1, were assembled from a variety of governmental and commercial sources. The listed covariates are a subset from a larger list and were included in the model partly based on their substantive, contextual, and empirical relationships with the 6 outcomes. In addition, to account for multiple years of data, appropriate numbers of dummy variables were also included as predictors. Thus the estimates derived from the model borrow strength across areas as well as time.

TABLE 1 ABOUT HERE

A hierarchical Bayesian approach was developed to obtain model-based estimates derived from three types of direct county-level estimates: (1) NHIS estimates for households with telephones; (2) NHIS estimates for households without telephones; and (3) BRFSS estimates (telephone). The differences between (1) and (2) provide information about errors due to noncoverage; and the differences between (1) and (3) provide information about nonresponse bias (although other factors such as mode and contextual effects might also be present), assuming that NHIS estimates are unbiased. Not all three direct estimates are available for every county. Since most of the 3,114 counties were included in the BRFSS sample, the BRFSS estimates are available for most of the counties. In contrast, among the roughly 25% of the counties included in the NHIS sample, most contained sampled households with telephones, whereas about 40% contained sampled households without telephones. Thus, NHIS telephone estimates are available for approximately 25% of U.S. counties, whereas NHIS non-telephone estimates are available for about 10% (40% of 25%) of U.S. counties.

1.4 Outline of the Paper

The rest of this paper is organized into four sections. Section 2 assesses the potential biases due to noncoverage and nonresponse by examining the distributions of the three types of direct estimates. The distributions of 1990 and 2000 telephone coverage rates are also assessed to study the potential impact of making estimates solely based on the BRFSS. These investigations set the stage for needing to combine information from the BRFSS and NHIS. Section 3 develops a hierarchical Bayesian model to combine information from the two surveys and describes the algorithm used to derive the combined county-level estimates. A Markov Chain Monte Carlo method is used to compute the posterior mean and standard deviation of population proportions for each year and county. Section 4 provides summaries of the distributions of combined estimates for 6 outcomes. More detailed analysis is given for two outcomes, current smoking among men and mammography. We also compare the combined estimates for these two outcomes to model-based estimates based solely on the BRFSS. The latter estimates were obtained based on a hierarchical model using just the direct estimates from the BRFSS and the same covariates listed in Table 1. Section 5 concludes with a discussion, including limitations of our modeling and estimation procedures and directions for future research.

2 Comparison of Direct Estimates

Both the NHIS and BRFSS employ complex survey designs involving weighting factors to adjust for unequal probabilities of selection, nonresponse, and post-stratification. Though streamlined in 1997, the BRFSS designs vary by state. In addition, the NHIS uses a multistage selection process. We therefore constructed weighted estimates based on the sample from each county and computed the design-based variance using Taylor's linearization approach (Binder 1983) for the specific designs in the NHIS and each state in the BRFSS. For county $j = 1, 2, \ldots, J$ and year $t = 1, 2, \ldots, T$, let p_{xjt}, p_{yjt}, and p_{zjt} denote the weighted prevalence estimates based on NHIS households with telephones, NHIS households without telephones, and the BRFSS, respectively. As mentioned in the preceding section, not all three estimates are available for every county and year. Each estimate is design-unbiased, and comparing these estimates may provide information about the extent of noncoverage and nonresponse bias. Nelson et al. (2003) compared NHIS and BRFSS national estimates for a number of outcomes using data from 1997 for both surveys. Here we compare the county level estimates.

We select two outcomes for a detailed investigation, current smoking for men aged 18 or older and mammography screening during the past 2 years for women aged 40 years or older. Figure 1 provides the means and standard deviations of the estimated county prevalence rates of current smoking among men for each year (that is, means and standard deviations of p_{xjt}, p_{yjt}, and p_{zjt} across the J counties for each t). It appears that the current-smoking rate for men living in households without telephones is almost twice the rate for those living in households with telephones. The distributions for the BRFSS and for the NHIS telephone households are only modestly different, with the largest apparent difference being for 1997. Thus, noncoverage bias, and perhaps to a lesser extent nonresponse bias, may be important issues for this outcome.

FIGURE 1 ABOUT HERE

Figure 2 gives the same information for mammography screening rates for the collapsed years 1997/1998 and 1999/2000. The collapsing was necessary because the mammography questions were not asked every year in both surveys. It appears that the mammography screening rates for the non-telephone households are about half the rates for the telephone households. The screening rates are similar between NHIS telephone households and BRFSS households.

FIGURE 2 ABOUT HERE

Similar patterns were observed for the remaining 4 outcomes, with some showing large differences between NHIS telephone households and BRFSS households. Table 2 gives national-level estimates, based on NHIS public-use files, of the prevalence rates of the 6 outcomes for telephone and non-telephone households. It is fairly obvious that noncoverage bias can be substantial in estimates that solely rely on telephone surveys.

TABLE 2 ABOUT HERE

To further study the possible impact of noncoverage bias, we computed the percentiles of the county-level telephone noncoverage rates based on the 1990 and 2000 censuses. The results for the two censuses are quite different. Based on the 1990 census, about 5.2% of the households in the nation did not have telephones, and the county-level rates ranged from 0.5% to 59.7%. Based on the 2000 census, however, the national percentage of households with no telephones was about 2.4%, and the county-level rates ranged from 0.4% to 46.1%. One might expect that the telephone coverage rates would increase between 1990 and 2000. However, an issue that could have contributed to the apparent increase involves the questions in the census about telephones in the house. In 1990, the relevant census question was, "Do you have a telephone in this house or apartment?" This question was similar to the one asked in the NHIS (1997-2000). In the 2000 census, however, the question was, "Is there telephone service available in this house, apartment, or mobile phone from which you can both make and receive calls?" This question was slightly different from those in the 1990 census and the 1997-2000 NHIS. The 2000 census question did not distinguish between cellular and land-based telephones, and it asked about service rather than the existence of a telephone in the house.

Regardless of which telephone rates are taken into consideration, given the range of county-level coverage rates, the extent of noncoverage bias can be substantial for some areas and modest for others. Combining information based on a model that reflects both noncoverage and nonresponse bias should improve the accuracy of county-level estimates. Despite the issue of question wording, we used 2000 census telephone coverage rates in developing combined estimates (as discussed later), as 2000 is closer in time to the survey years 1997-2000.

3 Model and Inference

3.1 Model

Hierarchical models for small area estimation have a long history, beginning with Fay and Herriot (1979). A Bayesian approach for small-area estimation was discussed in Dempster and Raghunathan (1985). Advances in computing have resulted in the ability to fit realistic, but complex, Bayesian models to obtain small-area estimates. See Rao (2003) for a comprehensive review of design-based, empirical Bayes, and Bayesian approaches.

We adopt a Bayesian approach, using a hierarchical model involving three stages. In the first stage, we develop an approximate sampling distribution for the three direct estimates (NHIS telephone, NHIS non-telephone, and BRFSS) conditional on county-level population parameters. Potential nonresponse and noncoverage errors are expressed in terms of differences in the expected values of the sampling distributions of the direct estimates. Complex sample designs are incorporated by using weighted estimates as direct estimates and by using design effects in computing the sampling variances and covariances. In the second stage, we model the between-county variation in the population parameters, incorporating county-level covariates. Finally, the modeling process concludes with a diffuse proper prior distribution for the unknown parameters in the second-stage model. In developing these models, we have taken a pragmatic approach by keeping the models relatively simple from a computational perspective, given the large number of counties, potential applications to a large number of outcomes, and a desire to develop a framework that could be used routinely for producing small-area estimates. Limitations of our approach, and possible alternatives, are discussed in Section 5.2.

3.1.1 Stage 1: Sampling Distribution

The first task is to approximate the sampling distribution of the direct estimates given that the sample designs and post-survey adjustment procedures differed between the BRFSS and the NHIS as well as within the BRFSS across the states and DC. We incorporate the complex survey design features by expressing the sampling variances in terms of the effective sample sizes for simple random samples (Kish 1995). Specifically, suppose that n_{rjt} and v_{rjt} are the sample size and estimated design-based variance, respectively, for the direct design-based estimate p_{rjt} (under whatever design was used by the survey), where $r = x, y, z$ and x, y, and z denote NHIS telephone, NHIS non-telephone and BRFSS (telephone) households. Suppose that P_{rjt} is the corresponding population proportion. Since the estimated variance based on a simple random sample would be $p_{rjt}(1 - p_{rjt})/n_{rjt}$, the estimated design effect is $d_{rjt} \equiv v_{rjt}/[p_{rjt}(1 - p_{rjt})/n_{rjt}]$. Thus, the estimated effective sample size is $\tilde{n}_{rjt} \equiv n_{rjt}/d_{rjt}$. The sampling distribution of p_{rjt} has the design-based expected value P_{rjt} and an approximate design-based sampling variance of $P_{rjt}(1 - P_{rjt})/\tilde{n}_{rjt}$. Thus, design effects provide a unified framework for dealing with differing designs and post-survey adjustments.

Since the sampling variances depend on the population proportions, we use the arcsine-square root transformations of the direct estimates as in Efron and Morris (1975) to approximately stabilize variances, which allows us to simplify our modeling and computation greatly. Our approximate sampling distribution for a direct estimate is

$$\sin^{-1}\sqrt{p_{rjt}} \sim N(\sin^{-1}\sqrt{P_{rjt}}, (4\tilde{n}_{rjt})^{-1}).$$

In developing this approximate sampling distribution, we treat the design effect as fixed at its estimate. This is similar to the customary practice of assuming that the sampling variance is known when using many hierarchical models for small-area estimation (Fay and Herriott 1979; Datta, Fay, and Ghosh 1991; Ghosh, Nangia, and Kim 1996; and Rao 1999).

In the practical implementation, we were unable to estimate design effects for some counties due to small sample sizes. In these cases, we imputed the average design effect of 1.3 for the NHIS and 1.1 for the BRFSS. If $\tilde{n}_{rjt}$ was less than 1, we reset it to 1.

In developing the joint sampling distribution for the direct NHIS estimates for telephone and non-telephone households and the direct BRFSS estimate, another issue to be addressed was the correlation between direct estimates for telephone and non-telephone households in the same county. Since the NHIS employs a multistage complex sample design, it is possible that telephone and non-telephone households may share the same primary and secondary sampling units. Though the resulting correlation may be small, especially if we condition on the county-level population proportions, we computed the design-based estimate of the correlation at the national level and incorporated it when specifying the joint sampling distribution of the direct estimates. Obviously, given the independent selections across the BRFSS and the NHIS, there will be no correlation between BRFSS and NHIS estimates conditional on county-level proportions.

Exploiting the variance stabilizing properties of the arcsine-square root transformation and the aforementioned correlation issues as well, we approximate the joint sampling distribution by a trivariate normal distribution,

$$\begin{pmatrix} x_{jt} \\ y_{jt} \\ z_{jt} \end{pmatrix} = \begin{pmatrix} sin^{-1}\sqrt{p_{xjt}} \\ sin^{-1}\sqrt{p_{yjt}} \\ sin^{-1}\sqrt{p_{zjt}} \end{pmatrix} \sim N_3 \left[\begin{pmatrix} \theta_{jt} \\ \phi_{jt} \\ (1+\delta_{jt})\theta_{jt} \end{pmatrix}, 4^{-1} \begin{bmatrix} \tilde{n}_{xjt}^{-1} & \rho_t(\tilde{n}_{xjt}\tilde{n}_{yjt})^{-1/2} & 0 \\ & \tilde{n}_{yjt}^{-1} & 0 \\ & & \tilde{n}_{zjt}^{-1} \end{bmatrix} \right],$$

where $\theta_{jt} = sin^{-1}\sqrt{P_{xjt}}$ is the arcsine-square root of the population proportion for those households equipped with telephones in county j at time t, $\phi_{jt} = sin^{-1}\sqrt{P_{yjt}}$ is for those households without telephones, δ_{jt} measures the proportionate bias in the BRFSS estimate relative to the NHIS estimate, and ρ_t is the correlation between the NHIS sample estimates from telephone and non-telephone households, which we fix at a value that is pre-estimated at the national level as mentioned above.

The manner in which BRFSS nonresponse bias enters into the above model, that is, by a factor of the form $1+\delta$ (with subscripts omitted for brevity) when proportions are on the transformed (arcsine-square root) scale, implies a bias of roughly the same form when proportions are on the original scale. This follows from the fact that, for a given δ and population proportion P within a reasonable range of values, we can find δ^* such that $(1+\delta)sin^{-1}\sqrt{P} \approx sin^{-1}\sqrt{(1+\delta^*)P}$. The population prevalence rates across all 6 outcomes are typically expected to be between 15% and 85%. In an empirical investigation of the adequacy of the approximation for $|\delta| \leq 0.2$ and $0.1 \leq P \leq 0.9$, the two sides of the approximate equality had an R-square of about 98.5%. Furthermore, a scatter plot of the NHIS direct estimates for telephone households and the BRFSS direct estimates showed a linear relationship, which further lends empirical support for this model. Hence, for a given estimate of P and δ, we can approximately estimate δ^* to assess bias on the original scale.

3.1.2 Stage 2: Between-Area Model

Let $\boldsymbol{\omega}_{jt} = (\theta_{jt}, \phi_{jt}, \delta_{jt})$ denote a 3×1 vector of county-level population parameters. Let $\boldsymbol{U}_{jt}$ be a $p \times 1$ vector of covariates for county j, including $T-1$ dummy variables corresponding to years and the intercept. In the second stage of the model, the county-level population parameters $\boldsymbol{\omega}_{jt}$ are assumed to follow a trivariate normal distribution,

$$\boldsymbol{\omega}_{jt} \sim N_3(\boldsymbol{\beta}\boldsymbol{U}_{jt}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\beta}$ is a $3 \times p$ matrix of regression coefficients and $\boldsymbol{\Sigma}$ is a 3×3 covariance matrix. However, the parameter space needs to be restricted so that all 3 population proportions lie between 0 and

1 (see the Appendix). The conditional and joint distributions are, therefore, truncated normal distributions. In the actual application, the 18 covariates other than the dummy variables were included in $\boldsymbol{U}$ on the logarithmic scale to reduce the impact of skewness and standardized to improve numerical stability.

3.1.3 Stage 3: Hyperprior

Finally, we assume a diffuse proper prior for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, with columns of $\boldsymbol{\beta}$ having independent multivariate normal distributions, $N_p(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o)$. In the practical implementation, we fixed $\boldsymbol{\beta}_o = \mathbf{0}$ and $\boldsymbol{\Sigma}_o = 10^4\mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix; for $\boldsymbol{\Sigma}$, we specified an inverse-Wishart distribution with $d_o = 4$ degrees of freedom and scale matrix $\mathbf{R}_o$ which we fixed at $10^{-4}\mathbf{I}_3$. These prior distributions are relatively diffuse but assure that the posterior distribution will be proper.

3.2 Inference

The ultimate objective is to obtain prevalence estimates (with standard errors) for all counties $j = 1, 2, \ldots, J$ and times (years) $t = 1, 2, \ldots, T$ (i.e., $T \times J$ small-area estimates). Suppose that M_{jt} denotes the proportion of target subjects living in households with telephones. Then the inferential quantity of interest is the composite proportion, $\mu_{jt} = M_{jt} sin^2\theta_{jt} + (1 - M_{jt}) sin^2\phi_{jt}$. We used telephone coverage rates from the 2000 census as estimates of M_{jt}. Using 1990 telephone coverage rates as estimates alters the inferences only modestly.

Given the complex nature of the model and the large number of estimands of interest, a simulation technique is the most computationally feasible method of estimation. Estimation via simulation is accomplished by drawing values from the posterior distribution of $\mu = \{\mu_{jt}, j = 1, 2, \ldots, J, t = 1, 2, \ldots, T\}$ given the observed data $D = \{x_{xobs}, y_{yobs}, z_{zobs}; \boldsymbol{U}_{jt}, j = 1, 2, \ldots, J, t = 1, 2, \ldots, T\}$, where *xobs*, *yobs*, and *zobs* are sets of county identifiers for which the respective direct estimates x, y, and z are available. The Markov Chain Monte Carlo technique of Gibbs sampling (Gelfand and Smith 1991; Tierny 1991) provides a convenient framework to draw values from the joint posterior density of $\{\boldsymbol{\omega}_{jt}, j = 1, 2, \ldots, J, t = 1, 2, \ldots, T; \boldsymbol{\beta}; \boldsymbol{\Sigma}\}$ given D. The procedure involves drawing from the joint posterior distribution of very large number of parameters, $k = 3(TJ + p) + 6$. For example, for the smoking prevalence rates, $J = 3114$, $T = 4$, and $p = 24$ (20 covariates, intercept, and 3 dummy variables for years), and thus $k = 37,446$; for the cancer screening rates, $J = 3114$, $T = 2$, and $p = 22$ (20 covariates, intercept, and a dummy variable for years), and thus $k = 18,750$. However, the conditional distributions in the Gibbs sequence either involve normal or inverse-Wishart distributions, so that creating draws is relatively straightforward. Details of the specific conditional distributions for this model are given in the Appendix.

For each drawn value of θ_{jt} and ϕ_{jt}, a draw of $\mu_{jt} = M_{jt} sin^2\theta_{jt} + (1 - M_{jt}) sin^2\phi_{jt}$ can be computed. The draws of μ_{jt} can be used to approximate the posterior distributions of the estimands of interest. The posterior mean and variance of μ_{jt} can be computed using all draws in the Gibbs sequence after ignoring a sufficiently large initial number of draws. For sufficiently large n, the set of draws that includes every n^{th} draw in the Gibbs sequence can be treated as approximately independent draws after the convergence criterion has been met.

4 Descriptive Analyses of Estimates

In our application, 10 parallel sequences, each of length 10,000, were used in Gibbs sampling. The first 5,000 draws from each sequence were discarded, and then the next 5,000 were included in computing posterior means and variances. Draws were pooled across the 10 parallel sequences, so that a total of 50,000 draws were used in computing each summary.

The Gelman-Rubin statistic R (Gelman and Rubin, 1992) was used to assess convergence for each μ_{jt}, $j = 1, 2, \ldots, J, t = 1, 2, \ldots, T$, as well as $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Across the six outcomes, the largest value of R was 1.053 for a proportion (μ_{jt}) and 1.070 for a regression parameter (component of $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$).

All programs were developed using the GAUSS programming language (Aptech Systems 2003). Computations were performed using a Dell Optiplex GX400 computer with a 1.7 GHz Intel Pentium 4 processor and 1 GB of internal memory. As examples of computing time, the estimation for current smoking among men, for which there are direct estimates for four time periods (1997, 1998, 1999, and 2000), took roughly 25 hours; and the estimation for mammography, for which there are direct estimates for two time periods (1997/1998 and 1999/2000), took roughly 12 hours.

Most of our descriptive analyses are based on the posterior means of the county population rates $\mu_{jt}, j = 1, 2, \ldots, J, t = 1, 2, \ldots, T$. However, since the draws from the posterior distribution of μ_{jt} are obtainable, more refined analyses are possible.

4.1 Analysis of Current-Smoking and Mammography Screening Rates

Figure 3 gives a color-coded chloropleth county-level map based on the combined NHIS/BRFSS estimates of current-smoking rates for men in 2000, with the rates grouped into 6 categories. Several counties in Kentucky, Missouri, Pennsylvania, the Virginias, and the Carolinas have high estimated prevalence rates of smoking among men. Figure 4 gives the corresponding map for mammography screening rates based on 6 categories. The screening rates are typically lower in the counties that have higher smoking prevalence.

FIGURES 3 AND 4 ABOUT HERE

Such maps would be useful for policy makers and researchers in identifying counties with high or low smoking or screening rates. For example, further understanding of local policies in areas with low smoking rates or high screening rates may provide information about potential intervention strategies. The counties with high smoking rates or low screening rates may be target areas of opportunity for interventions and allocation of resources.

Table 3 gives the descriptive statistics for 4 smoking rates, $(current, ever) \times (Male, Female)$ for each year. Table 4 gives the same information for the two cancer screening rates. These tables show that the 10^{th} and 90^{th} percentiles are between 15% and 85% for all outcomes and all years, which provides support for the linear approximation discussed in Section 3.1.1.

TABLES 3 and 4 ABOUT HERE

4.2 Comparison with Estimates Based on BRFSS

Given confidentiality concerns and limitations on data sharing, a comparison of the NHIS/BRFSS combined estimates with estimates based solely on BRFSS data could be informative. The BRFSS data are more widely available, with area identifiers suitable for county-level estimation.

We used a hierarchical model similar to the one used in Efron and Morris (1975) to derive the BRFSS-alone estimator, with

$$sin^{-1}(\sqrt{p_{zjt}}) \sim N(\mu_{jt}, 1/(4\tilde{n}_{zjt}))$$

as the first-stage model and

$$\mu_{jt} \sim N(\boldsymbol{\gamma U}_{jt}, \sigma^2)$$

as the second-stage model; these models are analogous to those described in Sections 3.1.1 and 3.1.2 for the combined NHIS/BRFSS estimates. Using a prior distribution similar to that in

Section 3.1.3, with $\boldsymbol{\gamma} \sim N_p(\mathbf{0}, 10^4\mathbf{I}_p)$ and σ^2 following an inverse-chi-square distribution with 2 degrees of freedom and scale parameter 10^{-4}, we obtained draws from the posterior distribution of $sin^2(\mu_{jt})$ via Gibbs sampling. The posterior means and standard deviations were computed from these draws for comparison with the combined NHIS/BRFSS estimates. The difference in the point estimates, $\widehat{\mu}_{BRFSS+NHIS} - \widehat{\mu}_{BRFSS}$, may provide information on the effects of the adjustments for noncoverage and nonresponse biases.

Figure 5 gives two histograms. The first histogram displays the differences in the estimates of current-smoking rates for men in 2000, and the second displays the differences in the estimated 1999/2000 mammography screening rates for all counties. The predominance of positive differences in the first histogram indicates that the current-smoking rates may be underestimated by using data from the BRFSS alone. Similarly, the predominance of negative values in the second histogram suggests that mammography screening rates may be overestimated by using data from the BRFSS alone.

FIGURE 5 ABOUT HERE

The apparent underestimation of smoking rates and the overestimation of mammography rates using the BRFSS alone are consistent with the finding that in general, telephone surveys such as BRFSS contain too few responders with low socioeconomic status (Goyder et al. 2002). Given the differences in the smoking and mammography rates between telephone and non-telephone households, and the variation in the county-level telephone coverage rates (and their correlation with socioeconomic factors), the combined estimation procedure attempts to provide compromise estimates.

5 Discussion

County-level estimates of the prevalence rates of cancer risk factors and screening are needed by cancer surveillance researchers as well as policy makers. Information from two popular surveys, both with advantages and disadvantages, has been combined to obtain county-level estimates for 6 outcomes. This project represents a collaborative effort of the National Cancer Institute, the National Center for Health Statistics (and its parent organization, the Centers for Disease Control and Prevention), and the University of Michigan. The empirical investigations have suggested that the combined estimation procedure helps to address noncoverage and nonresponse issues. It is possible to use this strategy to estimate prevalence rates of other factors such as obesity, dietary habits, and other cancer-related outcomes. The same methodology can be applied in other contexts as well, where there are multiple sources of data.

5.1 A Simulation Study

Our approach was evaluated using a simulation study as described in Xie (2004). The objective of the simulation study was two-fold: first, to check whether the computer code used to generate the estimates in the application was correct; and second, to check whether the Bayes estimates had desirable repeated-sampling properties. Using the parameter estimates from the analysis of current smoking for men, we generated 500 sets of pseudo-data from the NHIS and the BRFSS for 184 counties in Massachusetts, Michigan, and Minnesota. The first principal component based on 18 county-level covariates from the application (excluding two that were added near the end of the project) was used as a single covariate in the simulation, and only one year of data was generated. Thus, the simulated data sets may be viewed as simpler versions of the actual data used in the application. This simplification was necessary to complete the simulation study in a reasonable amount of time.

For each simulated data set, we obtained two sets of hierarchical Bayes estimates, one based on combined information from the NHIS and BRFSS data sets, and the other based on only the BRFSS data set. The combined estimates were practically unbiased and had good coverage properties, whereas the BRFSS-alone estimates were more biased and had worse coverage properties.

5.2 Limitations of Our Modeling and Estimation Procedures

5.2.1 Use of the Arcsine-Square Root Transformation

As was discussed in Section 3.1.1, we applied the arcsine-square root transformation to the outcomes in our model, and we exploited the variance stabilizing properties of this transformation to simplify our modeling and computation. Since the derivation of the arcsine-square root transformation for variance stabilization is rooted in large-sample theory, our model might be somewhat deficient, especially for counties with small sample sizes.

To investigate this issue in a simpler situation that does not involve combining information from two surveys, we compared the BRFSS-alone estimates of Section 4.2, which also incorporated the arcsine-square root transformation, with estimates obtained for a logistic/normal random effects model fitted to the BRFSS data using SAS PROC NLMIXED, for current smoking among men. (The latter approach cannot be applied to our problem of producing combined estimates in a straightforward manner.) The two procedures produced similar results, but the BRFSS-alone estimates that incorporated the arcsine-square root transformation tended to be smaller for counties with low smoking rates and/or small sample sizes. Although we do not know which estimates are preferable (since the truth is unavailable), the findings suggest some possible deficiencies in using the arcsine-square root transformation for small-sample situations.

Anscombe (1948), Freeman and Tukey (1950), and Mosteller and Youtz (1961) proposed alternative transformations for small-sample situations. We experimented with these transformations for our application. While use of the alternative transformations appeared to change the estimates somewhat, particularly for counties with small sample sizes, the convergence properties of the Gibbs sampling algorithm were substantially worse than when the usual arcsine-square root transformation was used.

Alternatives that would not rely on the arcsine-square root transformation or normality of the sampling distribution of the direct estimates are possible. For example, generalized linear models with binomial or Poisson errors, as suggested in Ghosh et al. (1998) and Farrell (2000), could be used but would complicate the modeling (e.g., in incorporating complex design features) and computational tasks greatly. For example, use of the logistic-normal model would require rejection sampling techniques such as the Metropolis-Hastings algorithm within the Gibbs sampler for over 37,000 parameters, and the resulting increased computing time needed would likely render the estimation procedure infeasible. An additional complication with this approach would be difficulties in incorporating correlations between the NHIS telephone and non-telephone estimates (but see also Section 5.2.2 below).

Hopefully, further research as well as increased computer power may enable us to avoid making simplifying assumptions. For the current project, however, we chose to incorporate such simplifying assumptions so that results could be feasibly obtained.

5.2.2 Pre-Estimation of Correlations and Design Effects

As was also discussed in Section 3.1.1, we chose to estimate the design effects d_{rjt} for the county-level prevalence rates and the correlations ρ_t between the NHIS telephone and non-telephone estimates and then treat them as fixed inputs into our Bayesian procedure. Although this is similar in spirit to treating sampling variances as fixed at estimates, which is common in small-area estimation (Fay and Herriott 1979; Datta, Fay, and Ghosh 1991; Ghosh, Nangia, and Kim 1996; and Rao 1999), it could result in underestimation of variability in the Bayesian procedure.

Comments similar to those in Section 5.2.1 could be made about alternatives to our simple pre-estimation procedures. For example, modeling the design effects as random would again require modeling the sampling variances as functions of the means, necessitating the use of a computational procedure such as the Metropolis-Hastings algorithm within the Gibbs sampler and increasing computing time substantially.

With regard to the correlations ρ_t, a possible alternative, albeit stronger, assumption would be to assume independence between the NHIS telephone and non-telephone estimates. The pre-estimated values of ρ_t in our project were all quite close to zero, so an assumption of independence might be reasonable. Such an assumption would simplify modeling, either in the context of our approach or with an approach that incorporates a generalized linear model (see Section 5.2.1), although the latter type of approach would still be much more computationally intensive.

5.2.3 Spatial Effects

In a problem such as ours that involves geographical units such as counties, it is natural to ask whether spatial modeling might improve the estimation for small areas. The model that we have used already incorporates correlations between parameters within counties (see Section 3.1.2), and more than a dozen county-level covariates have been included in our model to account for correlations between counties.

A spatial component could be added to the between-area model in Section 3.1.2 to account for omitted covariates. Rao (2003, Section 9.5) provides an excellent summary of applications of such methods to small-area estimation, the most successful of which have been to modeling population (mortality) data as opposed to sample survey data. As with the alternatives discussed in Sections 5.2.1 and 5.2.2, introduction of a spatial component would render the computational procedures substantially more complicated.

5.3 Possible Extensions

In the present paper, we considered only two data sources. The framework can be extended when three or more sources are available. Our approach can also be used to compute estimates for subdomains such as subpopulations based on region, age, race, gender, education, poverty, etc. These estimates may be obtained using publicly available BRFSS and NHIS data if county identifiers are not needed for subdomain estimation.

The modeled nonresponse biases in the BRFSS estimates (that is, the δs), did not explicitly involve county-level response rates. A sensible modification of the model would be to explicitly incorporate the county-level response rates, given the notion that the nonresponse bias in BRFSS estimates is smaller for counties with high response rates and larger for counties with low response rates. The county-level response rates were not available for this project. Currently, information on these rates is being collected, and a modified model explicitly incorporating the estimated rates is under investigation.

We also assumed that the residual covariance matrix $\mathbf{\Sigma}$ in the second-stage model is the same for all of the areas. It is conceivable that there may be more variation in the smaller counties than in the larger counties. Thus, the residual covariance matrix may depend on the population size of the county. A simple fix to this problem would be to stratify based on the size of the counties. A preliminary investigation suggests that this would lead to practically similar results. An alternative would be to explicitly model the residual covariance matrix as a function of covariates. Such modifications would make the modeling and computational tasks quite complex, and they may not be practical when several thousand county estimates of a large number of outcomes are needed on a production schedule.

Finally, the county-level estimates were obtained separately for each outcome. Another possibility would be to treat this as a multivariate problem, because one would expect these outcomes to be correlated with each other. Again, the complexity in the modeling and computational tasks may outweigh any benefits derived from any of the modifications suggested above.

Appendix

The Markov Chain Monte Carlo method used in developing county-level estimates in this article was based on Gibbs sampling using the following conditional distributions. For brevity, we omit all of the subscripts for θ, ϕ and δ. All of the sums are over J counties. We also let $a = \tilde{n}_x^{-1}/4$, $b = \tilde{n}_x^{-1/2}\tilde{n}_y^{-1/2}/4$, $c = \tilde{n}_y^{-1}/4$ and $d = \tilde{n}_z^{-1}/4$.

1. **Conditional distribution of $\boldsymbol{\beta}$**:

   Here $\boldsymbol{\beta}$ is a $3 \times p$ matrix of regression coefficients. Let $\widehat{\boldsymbol{\beta}} = \left(\sum_j \sum_t \boldsymbol{\omega}_{jt}\mathbf{U}_{jt}^T\right)\left(\sum_j \mathbf{U}_{jt}\mathbf{U}_{jt}^T\right)^{-1})$ be the least squares estimate. Denoting $\mathbf{\Sigma}_{\boldsymbol{\beta}} = \mathbf{\Sigma} \otimes \left(\sum_j \mathbf{U}_{jt}\mathbf{U}_{jt}^T\right)^{-1}$ and $\mathbf{\Sigma}_o^* = \mathbf{\Sigma}_o \otimes I_3$, it is straightforward to show that

   $$vec(\boldsymbol{\beta})|\, \theta, \phi, \delta, \mathbf{\Sigma}, Data \sim N\left((\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} + {\mathbf{\Sigma}_o^*}^{-1})^{-1}\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} vec(\widehat{\boldsymbol{\beta}}), (\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} + {\mathbf{\Sigma}_o^*}^{-1})^{-1}\right),$$

   where vec denotes a vector created from the columns of a matrix. When $\mathbf{\Sigma}_o = a_o\mathbf{I}_p$,the above multivariate normal distribution simplifies to

   $$N(a_o(a_o\mathbf{I}_{3p} + \mathbf{\Sigma}_{\boldsymbol{\beta}})^{-1} vec(\widehat{\boldsymbol{\beta}}), a_o(a_o\mathbf{I}_{3p} + \mathbf{\Sigma}_{\boldsymbol{\beta}})^{-1}\mathbf{\Sigma}_{\boldsymbol{\beta}}).$$

2. **Conditional distribution of $\mathbf{\Sigma}$**:

   Let $\mathbf{S} = \sum_j \sum_t (\boldsymbol{\omega}_{jt} - \boldsymbol{\beta}\mathbf{U}_{jt})(\boldsymbol{\omega}_{jt} - \boldsymbol{\beta}\mathbf{U}_{jt})^T/JT$. It is easy to show that

   $$\mathbf{\Sigma}|\boldsymbol{\omega}, \boldsymbol{\beta}, Data \sim Inverse - Wishart(JT + d_o, \mathbf{R}_o + JT\mathbf{S})$$

3. **Conditional distribution of $(\theta, \phi)^T$**:

   Let $z^* = z(1+\delta)^{-1}$; then we have

   $$\begin{pmatrix} (x+z^*)/2 \\ y \end{pmatrix} |\theta, \phi, \delta \sim N\left[\begin{pmatrix} \theta \\ \phi \end{pmatrix}, \begin{pmatrix} (a + d(1+\delta)^{-2})/4 & \rho b/2 \\ \rho b/2 & c \end{pmatrix}\right]. \qquad (1)$$

   It is straightforward to derive

   $$\left.\begin{pmatrix} \theta \\ \phi \end{pmatrix}\right| \delta, \boldsymbol{\beta}, \mathbf{\Sigma}, x, z, y$$

   as bivariate normal using the standard multivariate normal theory (Anderson, 1984) and the following Lemma in Lindley and Smith (1972).

Lemma: If $y|\theta_1 \sim N(A_1\theta_1, C_1)$ and $\theta_1|\theta_2 \sim N(A_2\theta_2, C_2)$ then $\theta_1|\theta_2, y \sim N(Bb, B)$ where $B = A_1^T C_1^{-1} A_1 + C_2^{-1}$ and $b = A_1^T C_1^{-1} y + C_2^{-1} A_2\theta_2$.

While constructing the conditional distribution, one needs to keep track of the pattern of missing data in the direct estimates.

(a) When only x is missing, equation (1) is replaced by

$$\begin{pmatrix} z^* \\ y \end{pmatrix} |\theta, \phi, \delta \sim N\left[\begin{pmatrix} \theta \\ \phi \end{pmatrix}, \begin{pmatrix} d(1+\delta)^{-2} & 0 \\ 0 & c \end{pmatrix}\right], \tag{2}$$

and the others are the same.

(b) When only y is missing, θ and ϕ can be drawn independently. The conditional distribution of θ can be derived from

$$(x + z^*)|\,\theta, \phi, \delta, \rho \sim N(2\theta, a + d(1+\delta)^{-2})$$

and $\theta|\,\phi, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$. The conditional distribution of ϕ is simply $\phi|\,\theta, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

(c) When only z is missing, equation (1) is replaced by

$$\begin{pmatrix} x \\ y \end{pmatrix} |\theta, \phi\delta \sim N\left[\begin{pmatrix} \theta \\ \phi \end{pmatrix}, \begin{pmatrix} a & \rho b \\ \rho b & c \end{pmatrix}\right].$$

(d) When both x and y are missing, θ and ϕ can be drawn independently. The conditional distribution of θ can be derived from

$$z^*|\,\theta, \phi, \delta, \rho \sim N\left(\theta, d(1+\delta)^{-2}\right)$$

and

$$\theta|\,\phi, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}.$$

The conditional distribution of ϕ is simply the conditional distribution derived from the second stage model: $\phi|\,\theta, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

(e) When both y and z are missing, θ and ϕ can be drawn independently. The conditional distribution of θ can be derived from $x|\,\theta, \phi, \delta, \rho \sim N(\theta, a)$ and $\theta|\,\phi, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$. The conditional distribution of ϕ is simply $\phi|\,\theta, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

(f) When both x and z are missing, θ and ϕ can be drawn independently. The conditional distribution of ϕ can be derived from $y|\theta, \phi, \delta \sim N(\phi, c)$ and $\phi|\theta, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$. The conditional distribution of θ is appropriate conditional normal, $\theta|\phi, \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

(g) When x, y and z are all missing, we simply draw θ and ϕ from

$$\left.\begin{pmatrix} \theta \\ \phi \end{pmatrix}\right| \delta, \boldsymbol{\beta}, \boldsymbol{\Sigma}.$$

4. **Conditional distribution of** δ:

   It is straightforward to see that the conditional distribution can be obtained from

   $$(z/d-1)|\, \theta, \phi, \delta, \rho \sim N\left(\delta, d\theta^{-2}\right)$$

   and $\delta|\, \theta, \phi, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

   When z is missing, and no matter whether x or y is missing, we simply draw δ from $\delta|\, \theta, \phi, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

   To keep all the draws of proportions within the range [0,1], the range of plausible values of θ, ϕ and δ, are $[0, \pi/2], [0, \pi/2]$ and $[-1, \pi/(2\theta)-1]$ respectively. Thus, the draws from all the these univariate normal distributions incorporated these restrictions.

The estimand of interest is μ, which is defined as $M\sin^2(\theta) + (1-M)\sin^2(\phi)$ where M is the proportion of telephone-equipped households in a given area.

Table 1: Outcomes and county-level covariates

Six Outcomes	Twenty Covariates
Current smoking (Men, Age ≥ 18) Current smoking (Women, Age ≥ 18) Ever smoked (Men, Age ≥ 18) Ever smoked (Women, Age ≥ 18) Mammogram in the past 2 years (Age ≥ 40) Pap smear test in the past 3 years (Age ≥ 18)	Percent Black in 1996 Percent Hispanic in 1996 Percent completed high school among persons 25 years and over in 1990 Percent completed college among persons 25 years and over in 1990 Percent Social Security benefit recipients in 1996 Percent below poverty in 1993 Per capita reported serious crimes in 1995 Civilian labor force unemployment rate in 1996 Per Capita social service establishments in 1995 Per capita wages and salaries adjusted for cost of living in 1996 Per capita property taxes in 1992 Per capita expenditures and obligations in 1997 Monday-Friday newspaper readership rate in 1997 Population per square mile in 2000 Buying power index in 2000 Median effective buying income index in 2000 Per household total retail plus eating and drinking sales in 2000 Percent blue collar workers in 2000 Two dummy variables for whether the county is from a large (population $>$ 1 million) metropolitan statistical area (MSA), a small (population $<$ 1 million) MSA, or a non-MSA

Table 2: National NHIS estimates (and their standard errors) of prevalence rates (in percents) for six outcomes for each year by household telephone status

Outcome	Telephone				No Telephone			
	1997	1998	1999	2000	1997	1998	1999	2000
Current smoking (Men)	26.5 (0.5)	25.4 (0.5)	24.6 (0.5)	24.9 (0.5)	51.1 (1.5)	49.9 (2.0)	53.5 (2.1)	49.8 (1.9)
Current smoking (Women)	21.3 (0.4)	21.2 (0.4)	20.8 (0.4)	20.5 (0.4)	43.9 (1.7)	44.2 (1.8)	44.9 (2.1)	42.6 (1.9)
Ever smoked (Men)	54.1 (0.5)	53.4 (0.5)	52.5 (0.6)	51.0 (0.5)	64.8 (1.7)	63.9 (2.5)	64.7 (2.4)	65.0 (2.3)
Ever smoked (Women)	40.7 (0.4)	40.3 (0.5)	40.3 (0.5)	39.6 (0.4)	53.1 (2.1)	52.6 (2.3)	52.4 (2.4)	51.9 (2.6)
	Cancer Screening							
	1997-1998		1999-2000		1997-1998		1999-2000	
Pap smear	79.3 (0.4)		80.8 (0.3)		74.2 (2.1)		74.6 (1.6)	
Mammography	67.4 (0.6)		67.4 (0.4)		40.9 (3.8)		33.4 (2.7)	

Table 3: Percentiles (unweighted) of the combined NHIS/BRFSS estimated percentages of current and ever smokers by gender and year in 1997-2000 across 3,114 counties

Outcome	Gender	Year	Percent				
			10	25	50	75	90
Current Smoking	Male	1997	21.3	24.3	27.3	30.2	32.6
		1998	20.2	23.1	26.1	29.0	33.0
		1999	18.8	21.6	24.6	27.5	30.0
		2000	19.1	22.1	25.1	27.8	30.3
	Female	1997	16.6	19.1	21.9	24.7	27.1
		1998	16.8	19.4	22.1	25.0	27.5
		1999	16.5	18.9	21.7	24.6	26.9
		2000	16.2	18.8	21.6	24.5	26.8
Ever Smoking	Male	1997	54.9	59.4	63.3	66.1	68.7
		1998	54.8	59.0	62.3	65.9	68.4
		1999	53.3	58.0	61.6	64.5	67.0
		2000	50.9	55.3	59.1	62.2	64.7
	Female	1997	31.2	35.4	39.8	44.0	46.9
		1998	31.8	35.8	40.3	44.4	47.4
		1999	31.8	36.1	40.4	44.5	47.5
		2000	31.1	35.4	39.7	43.8	47.0

Table 4: Percentiles (unweighted) of the combined NHIS/BRFSS estimated percentages of mammography and pap smear cancer screening rates in 1997-1998 and 1999-2000 across 3,114 counties

Screening Outcome	Percent	Year	
		1997-1998	1999-2000
Pap-smear	10	73.6	75.0
	25	75.8	77.2
	50	78.3	79.7
	75	81.0	82.3
	90	83.8	85.3
Mammography	10	54.1	53.1
	25	58.3	57.7
	50	62.9	62.1
	75	67.5	66.9
	90	71.9	71.2

Figure 1: Means and standard deviations of county-level direct estimates of current-smoking rates for men

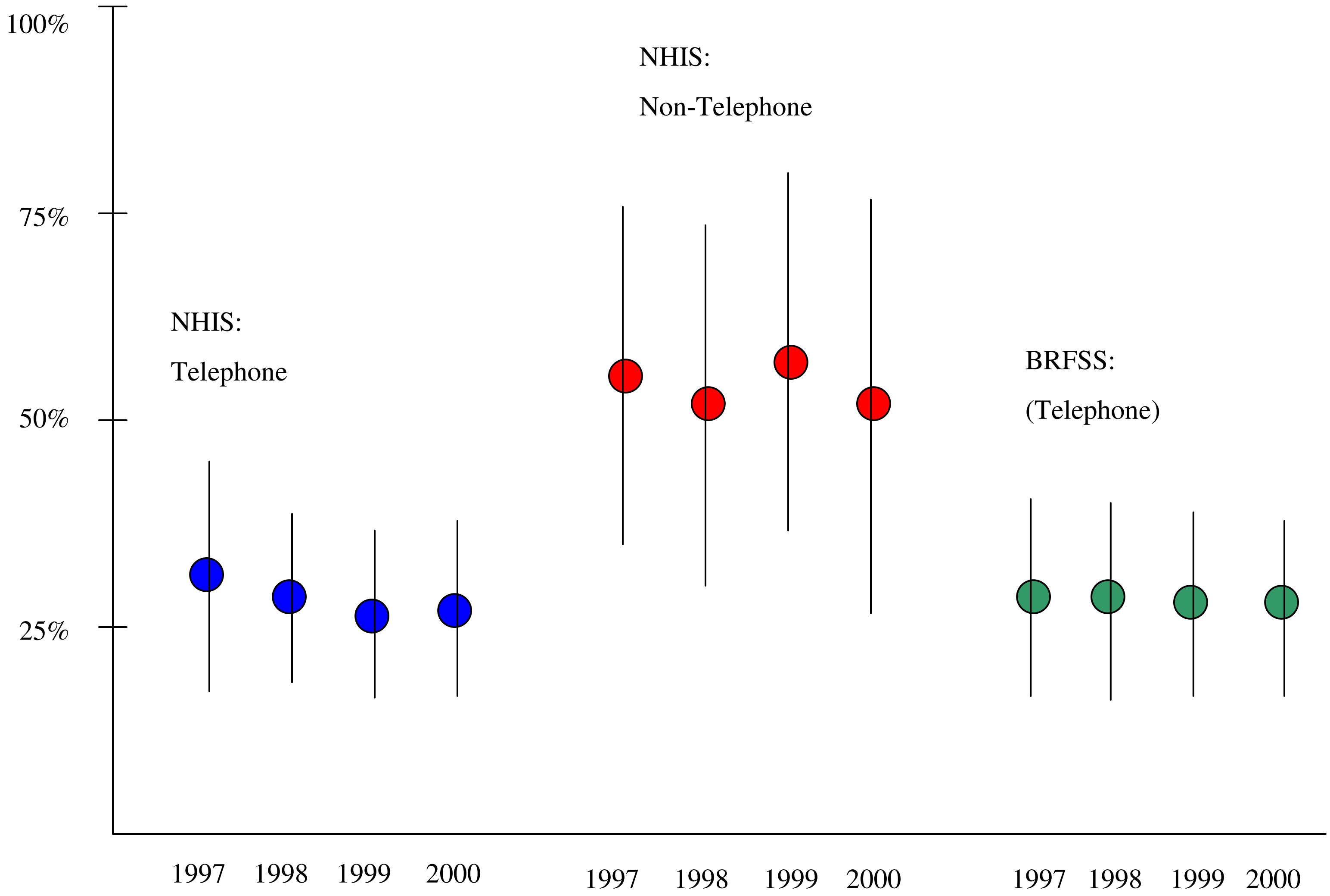


Figure 2: Means and standard deviations of county-level direct estimates of mammography rates

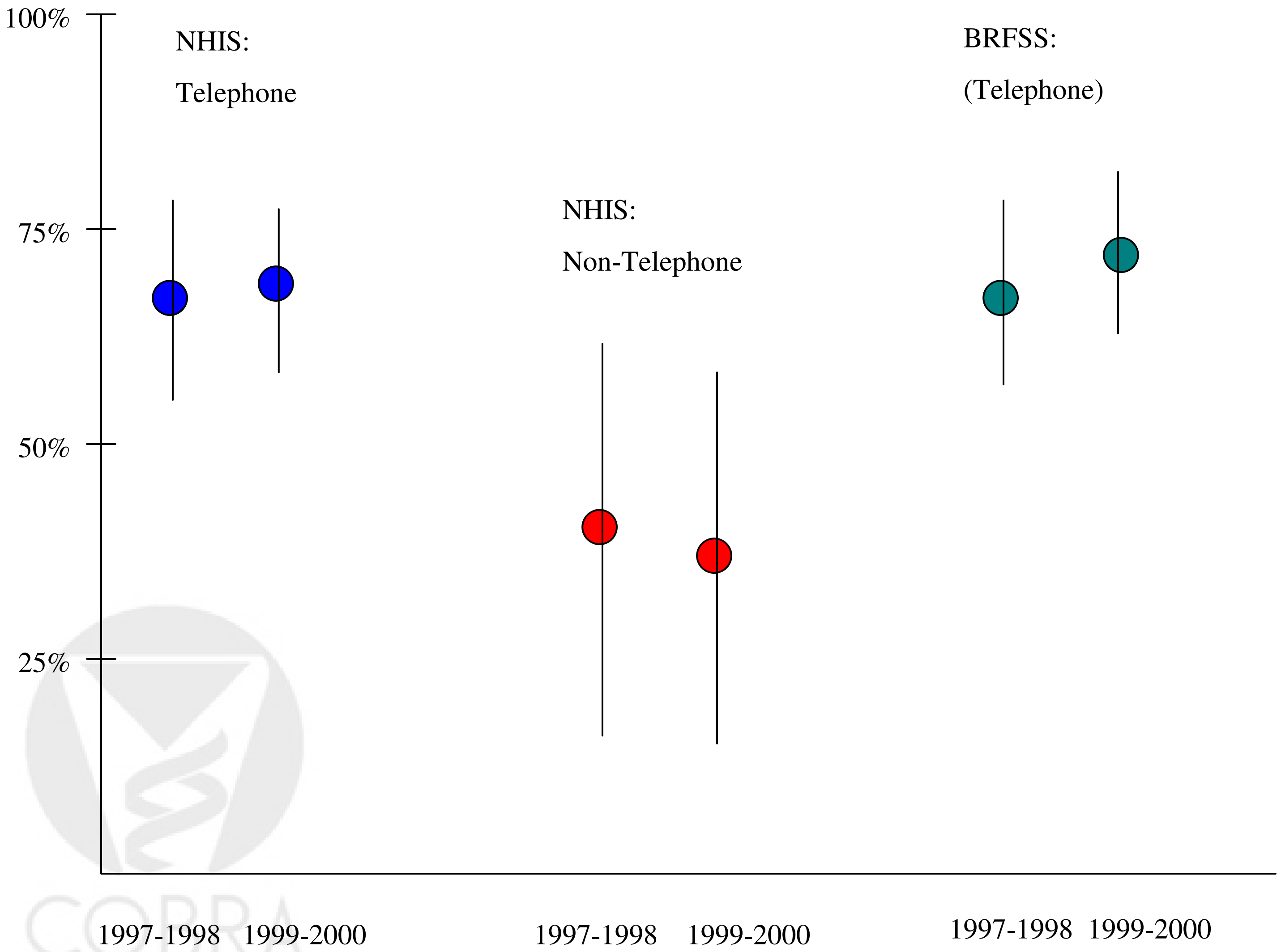


Figure 3: US map with county-level combined NHIS/BRFSS estimates of current-smoking rates for men in 2000

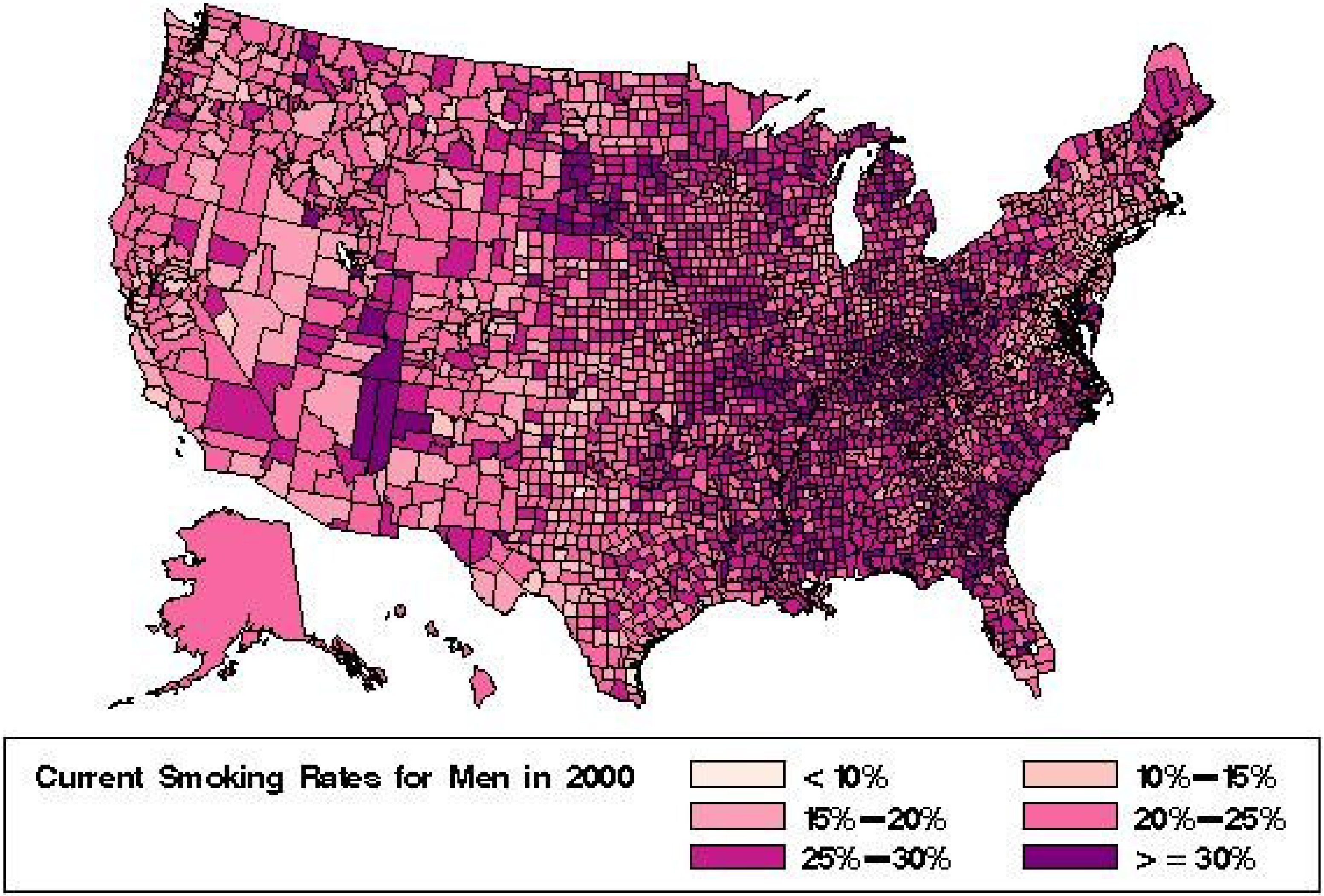


Figure 4: US map with county-level combined NHIS/BRFSS estimates of mammography screening rates in 1999-2000

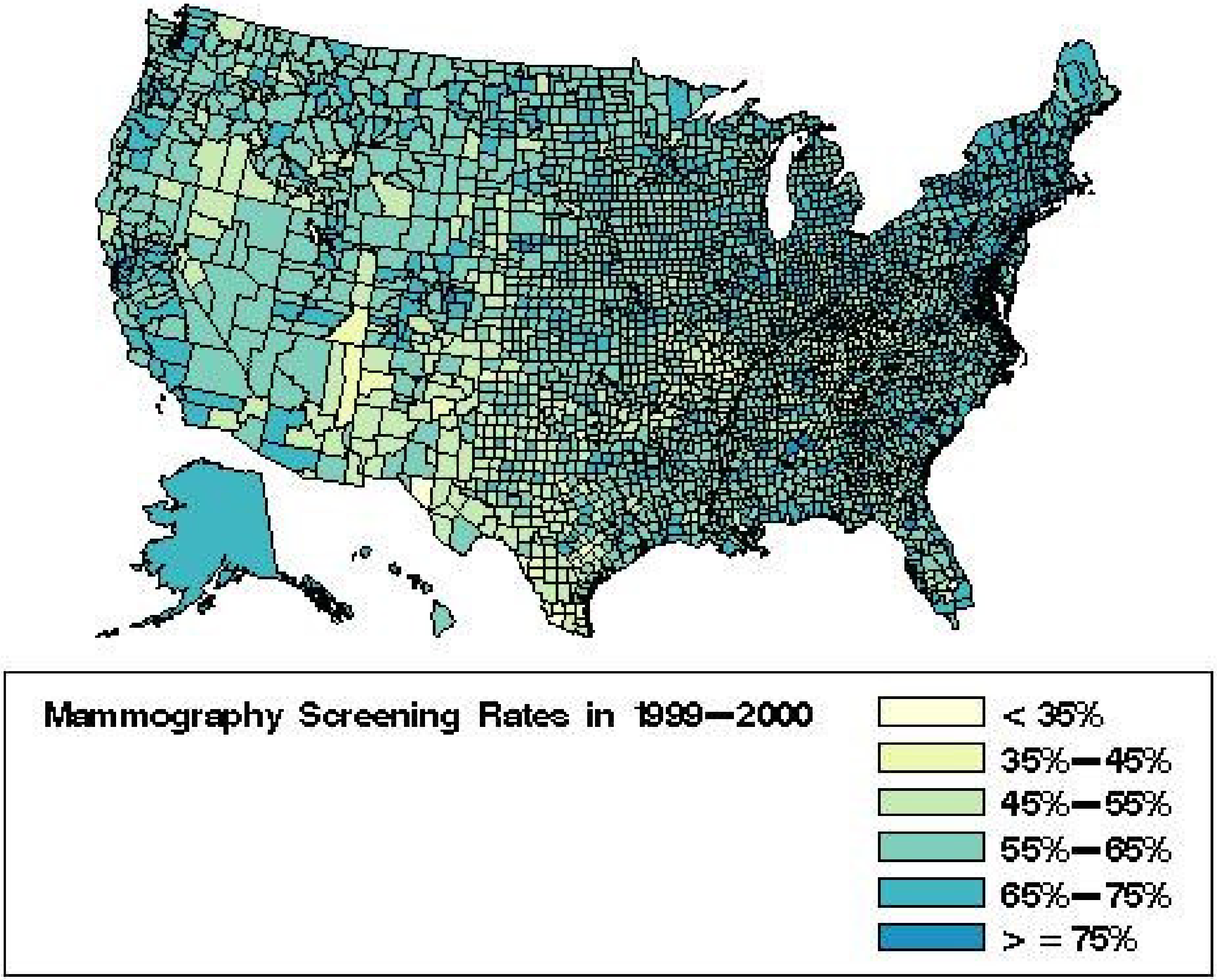
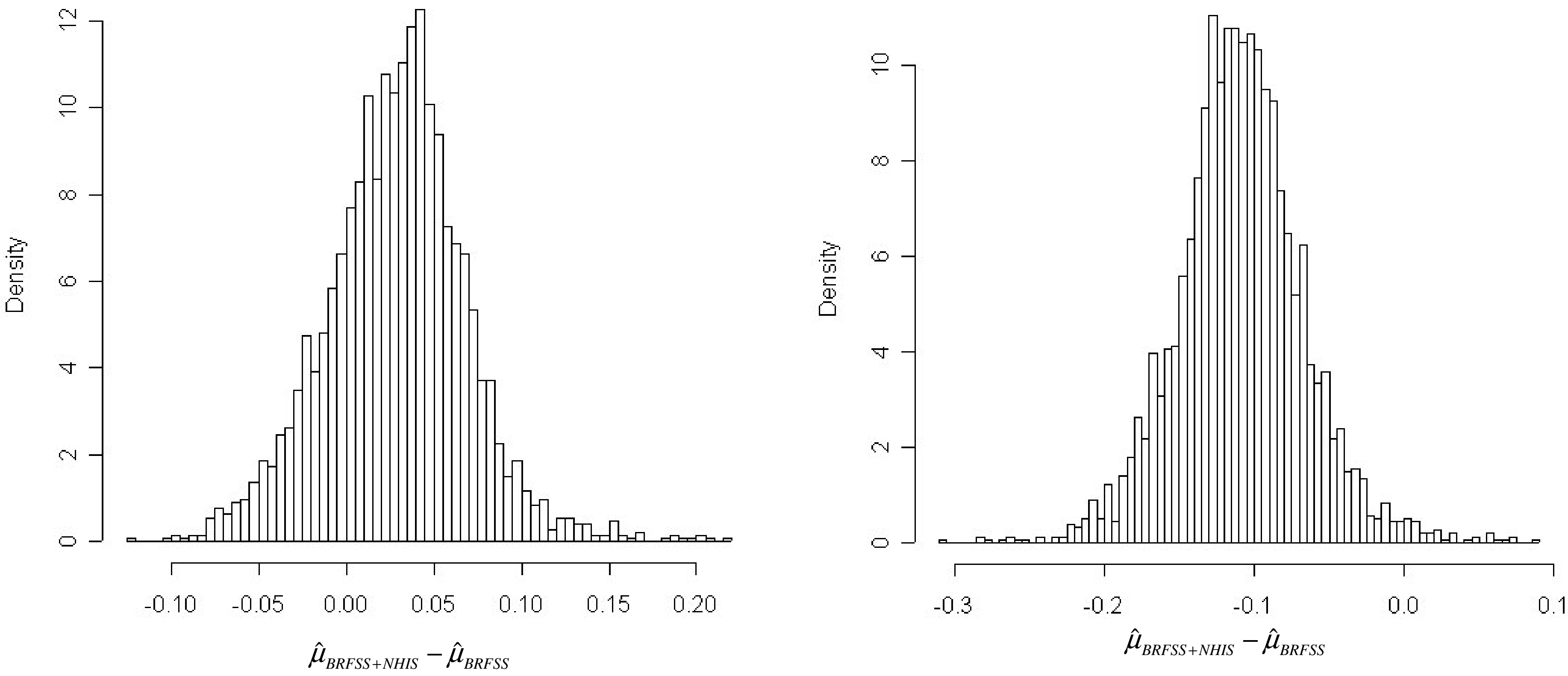


Figure 5: Histograms of the differences in combined (NHIS+BRFSS) and BRFSS alone estimates for two outcomes



Current-smoking rates for men: 2000

Mammography screening rates: 1999-2000

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Anderson, J. E., Nelson, D. E., and Wilson, R. W. (1998). Telephone coverage and measurement of health risk indicators data from the National Health Interview Survey. *American Journal of Public Health*, **88**, 1392-1395.

Anscombe, F.J. (1948) The transformation of Poisson, binomial and negative-binomial data. *Biometrika* 35,246-54

Aptech Systems (2003) Gauss: Advanced Mathematical and Statistical Systems, Version 5, Black Diamond, Washington.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.

Botman, S. L., Moore, T. F., Moriarty, C. L., and Parsons, V. L. (2000). Design and estimation for the National Health Interview Survey, 1995-2004. National Center for Health Statistics, *Vital and Health Statistics*, **2(130)**.

Breen, N., Wagener, D. K., Brown, M. L., Davis, W. W., and Ballard-Barbash, R. M. (2001). Progress in cancer screening over a decade: Results of cancer screening from the 1987, 1992 and 1998 National Health Interview Surveys. *Journal of the National Cancer Institute*, **93**, 23-32.

Centers for Disease Control and Prevention (2001), 2000 Summary Data Quality Report. Download from http://www.cdc.gov/brfss/technical_ infodata/quality.htm.

Datta, G.S., Fay, R. E. and Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation, in *Proceedings of the Bureau of the Census 1991 Annual research Conference*, U.S. Bureau of the Census, Washington DC, 63-79.

Dempster, A. P., and Raghunathan, T. E. (1985), Using a covariate for small area estimation: a common sense Bayesian approach. In *Small Area Statisitics: An International Symposium*, Platek et al., eds., New York: Wiley.

Efron, B. F., and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of American Statistical Association*, **70**, 311-319.

Elliott, M. R. and Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *Journal of Royal Statistical Society (Series C)*, 59, 595-609.

Farrell, P. J. (2000). Bayesian inference for small area proportions. *Sankhya, Series B*, 62, 402-416.

Fay, R. E., and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of American Statistical Association*, **74**, 269-277.

Freeman, M. F. and Tukey, J. W. (1950). Transformation related to angular and square root. *Annals of Mathematical Statistics*, 21, 607-11.

Gelfand, A. E., and Smith, A. M. F. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, **85**, 398-409.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Ghosh, M., Nangia, N., and Kim, D. (1996) Estimation of Median Income of four-person families: A Bayesian time series approach, *Journal of American Statistical Association*, 91, 1423-1431.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998), Generalized linear models for small area estimation, *Journal of American Statistical Association*, 93, 273-282.

Goyder, J.. Warriner, K. and Miller, S. (2002). Evaluating socio-economic status (SES) bias in survey non-response. *J. Off. Statist.*, 18, 1-12.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

Legler, J., Breen, N., Meissner, H., Malec, D., and Coyne, C. (2002). Predicting patterns of mammography use: a geographic perspective on national needs for intervention research. *Health Services Research*, **37**, 929-947.

Lepkowski, J. M. (1988). Telephone sampling methods in the United States. In *Telephone Survey Methodology*, Robert M Groves, et al., eds., New York: Wiley.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for linear models. *Journal of the Royal Statistical Society, Series B*, **34**, 1-41.

Mosteller, F. M. and Youtz, C. (1961). Tables of the Freeman-Tukey transformation for the binomial and Poisson distributions. *Biometrika*,48, 433-40

National Center for Health Statistics (2000a). 1997 National Health Interview Survey (NHIS) public use data release: NHIS survey description. Centers for Disease Control and Prevention, US Department of Health and Human Services, Hyattsville, MD.

National Center for Health Statistics (2000b). 1998 National Health Interview Survey (NHIS) public use data release: NHIS survey description. Centers for Disease Control and Prevention, US Department of Health and Human Services, Hyattsville, MD.

National Center for Health Statistics (2002a). 1999 National Health Interview Survey (NHIS) public use data release: NHIS survey description. Centers for Disease Control and Prevention, US Department of Health and Human Services, Hyattsville, MD.

National Center for Health Statistics (2002b). 2000 National Health Interview Survey (NHIS) public use data release: NHIS survey description. Centers for Disease Control and Prevention, US Department of Health and Human Services, Hyattsville, MD.

Nelson, D. E., Powell, E., Town, M. S., and Kovar, M. G. (2003). A comparison of national estimates from the National Health Interview Survey and the Behavioral Risk Factor Surveillance System. *American Journal of Public Health*, 93, 1335-1341.

Pickle, L. W., Feuer, E. J., and Edwards, B. (2001). Predicting cancer incidence in non SEER counties. *Proceedings of the Biometrics Section*, American Statistical Association, 45-52.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186..

Rao, J. N. K. (2003). *Small Area Estimation.* New York: Wiley.

SAS Institue. (2000). PROC NLMIXED, Online SAS Manual. Cary, N.C.

Swan, J., Breen, N., Coates, R. J., Rimer, B. K., and Lee, N. C. (2003). Progress in Cancer Screening Practices in the United States: Results from the 2000 National Health Interview Survey. *Cancer*, 97, 1528-1539.

Tierney, L. (1991). Exploring posterior distributions using Markov Chains. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 563-570.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of American Statistical Association*, **73**, 40-46.

Xie, D. (2004). *Combining Information from Multiple Surveys for Small Area Estimation.* Unpublished Ph.D. Thesis, Department of Biostatistics, University of Michigan, Ann Arbor.