AN INTERDISCIPLINARY CENTER ADDRESSING STATISTICAL ISSUES IN HEALTH RESEARCH AIMED AT HIGH-RISK POPULATIONS

William D. Kalsbeek, Robert P. Agans, Ashley F. Bowers, C. M. Suchindran, The University of North Carolina at Chapel Hill; and Ralph E. Folsom, Jr., Research Triangle Institute
William D. Kalsbeek, UNC-CH, 730 Airport Road, CB#2400, Chapel Hill, NC 27599-2400

KEY WORDS: Minority Health, Statistical Issues, Interdisciplinary Research, Health Research, CHSR

As long as population-based studies are an important element of the behavioral, policy, and surveillance sectors in health research, statistical issues accompanying the design, implementation, and analysis of these studies will remain an important challenge to the research community. This challenge can best be met by bringing together applied health researchers and statistical methodologists to address relevant problems in the field of public health. Recognizing the importance of interdisciplinary synergy, the School of Public Health at the University of North Carolina at Chapel Hill (UNC-CH), through the financial support of the National Center for Health Statistics, has established the *Center for Health Statistics Research (CHSR).*

The superordinate goal of the Center is to open up the channels of interdisciplinary communication in the field and to expand the research tools available to the public health community. The aim of this paper is to demonstrate how the CHSR attempts to meet this objective.

The CHSR

The CHSR was founded with the purpose of addressing statistical design and analysis issues tied to research focusing on minority and other populations groups known to be at relatively high risk to adverse health outcomes. Emphasis is given to methodological issues that arise in conjunction with existing substantive research efforts made by various organizations in the health research landscape of North Carolina.

The Center is composed of five core areas: an administrative core and four research areas.[1] The administrative core coordinates all Center activities and provides the infrastructure to facilitate: (i) the progression of current work; (ii) the exploration of new research directions; and (iii) the dissemination of all research findings. The four research areas serve as the intellectual nodes of the Center: Each separately addresses statistical and methodological issues as they pertain to various substantive matters in health research. Figure 1 presents the organizational structure of the CHSR and lists the organizational partners that contribute to the four major research investigations currently underway at the Center.

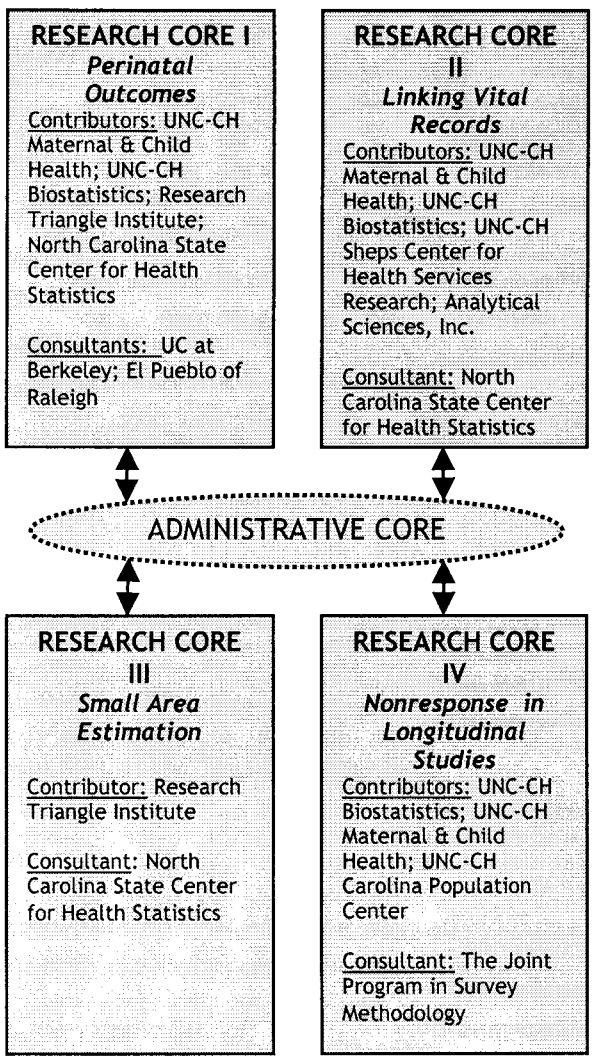

Figure 1. CHSR Structure

CHSR Research Areas

Our purpose in this section is to briefly sketch the work being done as part of the Center's current round of practice-oriented research. Each of the four areas described below examines statistical issues in current health research intended to promote health and prevent disease in high-risk populations.

[1] Information pertaining to the CHSR can be found on the Internet at the following address (http://www.sph.unc.edu/chsr/).

Research Core I: Perinatal Outcomes Research

This area brings together survey methodologists at the UNC-CH Department of Biostatistics with experience in sample design and questionnaire development, with UNC-CH Maternal and Child Health researchers investigating the perinatal health outcomes of Latino women. One interesting phenomenon that has arisen in the perinatal health literature is an apparent paradox among recent Mexican-American immigrants: The literature shows that these immigrants have relatively few low weight births (LBW) and infant deaths despite their low socioeconomic status and their limited use of prenatal services (Guendelman, 1998). Causes for lower rates of preterm delivery among Mexican-Americans are unknown at this time, but Buekens, Notzon, Kotelchuck, and Wilcox (2000) suspect that this phenomenon may be due partly to measurement errors in gauging gestational age.

The general aim of this study is twofold: (i) design researchers are applying prior research knowledge on sampling rare and elusive populations as well as employing dual-frame sampling methods, while (ii) the measurement team is applying cognitive interview techniques to reduce the potential for measurement error in assessing gestational age. The study's long-term goal is to bring together the sampling and measurement findings and make general recommendations for surveying the U.S. Latino population.

The Sampling Part. The main statistical novelty in the sampling part is the use of a theoretical framework for evaluating the error implications of sampling migrant seasonal farm workers (MSFWs), based on earlier work by Kalsbeek (1988). In this framework, the sample design being evaluated presumes that one must sample both in the spatial dimension (e.g., migrant camps and persons within camps) and the temporal dimension (i.e., to decide on which randomly chosen days during the period of study should select camps be visited). Since how one samples in the time and space dimensions impacts resulting survey error, one must formulate the error implications among plausible options and then compare the results.

Studying and improving methods to disproportionately sample high concentration areas and screening for Latino women living in the established residential population, will also be a focus of the sampling part of this research study. We intend to build on existing work on sampling rare populations (e.g., Kalsbeek and Cohen, 1978; Lepkowski, 1991) to further examine the statistical effectiveness and effects of oversampling area clusters and screening for targeted Latino women.

The Measurement Part. The measurement team is applying cognitive methods in the development of survey questions to measure gestational age. The cognitive interview is one qualitative method often used in questionnaire development and testing (see Forsyth & Lessler, 1991; Sudman, Bradburn, & Schwarz, 1996; Willis, Royston & Bercini, 1991). Cognitive interviewing is now widely used by the National Center for Health Statistics (NCHS), the Bureau of Labor Statistics (BLS), and the Bureau of the Census as well as major academic and private survey organizations.

Little work, however, has been done applying these cognitive methods to the development of survey instruments for immigrant populations, such as Mexican-Americans. According to Hines (1993), there are many barriers to overcome in designing surveys for immigrant populations: (i) problems with conceptual and linguistic equivalence; (ii) cultural sensitivity issues; and (iii) unfamiliarity with the use of survey measurement tools as well as nescience with the survey interview process.

Because our target population includes Spanish-speaking immigrants with little or no fluency in English, we are developing our measurement instrument in Spanish and face many of the measurement challenges outlined by Hines. Consequently, the issues of conceptual and linguistic equivalence are especially salient here: That some concepts may not have the same meaning or any meaning at all across cultures (conceptual equivalence) nor might they be uniformly understood by all respondents (linguistic equivalence), is something that we need to pay close attention to and document. We must also be cognizant of the fact that Mexicans typically add English words into their vocabulary and use a less formal Spanish than do other Hispanic groups (Elias-Olivares & Farr, 1991).

Another problem we are addressing involves dealing effectively with cultural sensitivity issues. Mexican-Americans, for example, tend to underreport sensitive acts, like sexual behavior, because of a strong sense of conventionality that pervades the culture. Furthermore, they may be more susceptible to social desirability effects (see McKay & Aguirre, 1994). Therefore, we are trying to come up with ways to reduce these potential biases when asking Latino women about their last menstrual period or other aspects of pregnancy that might be considered sensitive subjects for social discourse in surveys or interviews.

Finally, unfamiliarity with the use of survey measurement tools, like response scales and closed-end questions, can cause problems. Mexican respondents may be unpracticed in the skills required to complete a survey (e.g., choosing the most appropriate response options when provided with a list of alternatives). Another related problem is inexperience with the whole survey interview process. Aneshensel, Becerra, Fielder and Schuler (1989) remind us that different cultures tend to react differently to surveys. They found, for

example, that a general lack of experience with surveys resulted in greater attrition among Mexican-born participants. Therefore, we need to focus on and develop methods that encourage Mexican-American participation.

In sum, all of the above factors need to be taken into consideration when developing our cross-cultural instrument. The goal of the measurement team is not only to come up with reliable and valid measures of perinatal outcomes among Mexican-Americans, but also to make general recommendations for employing cognitive techniques to this population.

Taken together, it is hoped that the results of this study will shed light on the sampling and measurement barriers encountered when surveying Mexican-Americans, especially with regards to culturally sensitive topics. The findings should be of special interest to state and national health agencies responsible for assembling statistics on Latino immigrants (e.g., North Carolina State Center for Health Statistics, National Center for Health Statistics) as well as perinatal outcome specialists.

Research Core II: Linking Vital Records

This core area unites statistical modeling researchers at the UNC-CH Department of Biostatistics and a private research firm (Analytical Sciences, Inc.) with maternal outcome researchers at the UNC-CH Department of Maternal and Child Health. The focus of this area is to investigate statistical approaches to link birth records and health care data files to study maternal morbidity among Latino women and other racial and ethnic groups.

The objective of the linking process is to determine whether two or more records refer to the same person, object or event. Several methods are used in practice to complete the linking process: *Deterministic linking* is based on an exact match for all linking variables. *Probabilistic linking* is based on weighting the probabilities of agreement or disagreement among the linking variables in the files. *Staged linking* is based on isolating groups of records having one or more variables in common, and then using probabilistic techniques within these groups to effect linkage.

Modern computer technology has enhanced our capacity to conduct computer linkage of large public health data files. A key step in conducting this linkage is the development of an efficient *record-matching algorithm.* Formal development of a theory of record linkage started with the pioneering work of Fellegi and Sunter (1969). Jaro (1989) has extended their work and has developed computer software to implement probabilistic linkage of records. Additional insights to record linkage through statistical modeling techniques were provided by Copas and Hilton (1990). Except for these two papers, there has been little development of the statistical theory of record linking in the literature. However, recent advances in computer-intensive statistical methods, such as bootstrap techniques and classification methods, may be directly applicable to record linkage methodology.

Currently, there is widespread and growing use of linking multiple data sources in public health studies. The contexts in which these multiple linking of records is performed vary. For example, data from different cohorts are sometimes linked to avoid the cost of new surveys. Morbidity data, usually obtained through registries or hospital discharge records, are linked to mortality data to study the progression of disease and death. In longitudinal studies, files are linked to track cases over time. Linked files are sometimes used to construct sampling frames to examine the impact of multilevel characteristics on health outcomes or for program evaluation.

We are exploring the possibilities of applying these computer-aided statistical techniques to the task of probabilistic record linkage. Our research objectives include: (i) searching for the use of alternative statistical models in developing record matching algorithms for linking files; (ii) developing methods to assess the quality of the linkages; and (iii) exploring methods for adjusting outcome measures based on linked public health-administrative records.

Recently, we have been conducting a critical examination of the Copas and Hilton's "Hit and Miss" model. The model's assumptions, parameter estimations, fit of the model, and its generalizability to record linkage situations have been examined and preliminary results suggest that the model might not be generalizable beyond the specific applications presented in the Copas and Hilton (1990) paper. For example, the basic premise of the paper—that the two files to be linked contain equal number of records—may not be true for many situations.

We are also exploring the application of Classification and Regression Tree (CART) techniques to record linkage. Specifically, we have constructed a working data set from birth and infant death records, and are using this to experiment with various adaptations of CART. We have found one or two approaches that hold promise for record linkage, but we need to conduct further investigations before anything definite can be said about them.

Finally, in cooperation with the North Carolina Department of Health and Human Services and the Cecil G. Sheps Center for Health Services Research at UNC-CH, we are exploring the utility of linking North Carolina hospital discharge data with North Carolina birth certificate (Medicaid) files. It is hoped that this approach will provide better baseline levels of pregnancy-related morbidity. Currently, little is known

about the incidence and prevalence of pregnancy-related morbidity. This issue is especially salient because maternal mortality and low birth weight outcomes are higher in the U.S. than in many other industrialized countries.

Research Core III: Small Area Estimation

This area attempts to fuse research from the private sector, namely that being championed at the Research Triangle Institute (RTI), into work being done at the state health department level. The general goal is to develop methods for producing small area estimates based on prediction models utilizing census data, county data, and survey data. Our focus is on using a combination of federal and state data to produce estimates for sub-state areas. Identification of high-risk areas and population groups will help target intervention programs in the state of North Carolina.

Small area estimation (SAE) is the process of using statistical models to link national or state survey data outcome variables, like disease indicators, to local area predictors, like county demographic and SES variables, to predict local area disease prevalence rates. More powerful computer technology now makes it possible to use sophisticated hierarchical Bayes methods to fit the mixed Logistic and Poisson models that are ideally suited to SAE. Notable examples of this work include: (i) the SAE results of Malec, Sedransk, Moriarity and LeClere (1997) for binary outcomes from the National Health Interview Survey (NHIS); and (ii) the work of Nandram, Sedransk and Pickle (in press) reporting SAEs of mortality rates for U.S. Health Service Areas.

Our approach is to develop methods to generate useful SAEs from population-based samples and apply them to the public health sector. Our research involves not only developing new theoretical and practical SAE methods, but also demonstrating how these new methodologies can produce information that state and local health planners need to facilitate the effective development and targeting of health promotion and prevention programs, particular those aimed at population subgroups at great health risk (e.g., the obese, the physically inactive, etc.).

Our research objectives include: (i) finding ways to integrate comparable data from the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS) surveys to facilitate SAE for local health planning jurisdictions; (ii) developing dual-frame weight calibration methods for blending NHIS and BRFSS surveys to facilitate SAE; (iii) coordinating with North Carolina's State Center for Health Statistics in defining a set of ten to twelve outcome measures of interest to health planners in local areas of North Carolina for which SAEs are to be produced; and (iv) adapting RTI's new survey weighted SAE methodology to blended NHIS and BRFSS data to produce health statistics for counties and Health Service Areas (HSAs).

For our SAE study, we are using RTI's survey weighted version of current SAE methods based on the hierarchical Bayes (HB) approach for fitting logistic mixed models with random effects for the small areas. As applied to yes/no or binary type questionnaire outcomes where the statistics of interest are population rates of health risk, SAEs derived from these nonlinear logistic regression models behave like weighted averages or composites of the direct survey based prevalence estimate for the small area and the synthetic regression based predictor of the small area prevalence. This behavior, akin to that of a linear composite estimator, yields results that are close to the direct data based estimate when the local area has a relatively large sample. Otherwise, when the local area has a typically small sample, the SAE will shrink toward the synthetic logistic regression predictor that benefits from the full national sample to estimate its coefficients. The HB method provides a formal approach for specifying how far each small areas databased estimate will be shrunk in the direction of its regression predictor so that the mean squared error [variance + $(\text{bias})^2$] is minimized.

While this HB methodology is well developed for linear models, the logistic model results for binary outcomes are recent (Ghosh, Natarajan, Stroud & Carlin, 1998). Although these solutions lead to SAEs of prevalence that behave like linear composites, they are calibrated to reproduce the unweighted direct data prevalence estimator instead of the properly weighted survey proportion estimator. For a local area that has a relatively large disproportionately weighted sample, like the NHIS sample for Los Angeles County, California, the SAE based on an unweighted HB solution can be substantially biased. Our adaptation of the HB solution algorithm corrects this flaw in the available software such as BUGGS and MlwiN (Folsom, Shah & Vaish, 1999).

Research Core IV: Nonresponse Effects

This core area gathers a team of methodologists and health researchers (UNC-CH Biostatistics, UNC-CH Maternal & Child Health, Carolina Population Center) and takes a critical look at the effects of item and unit nonresponse in the National Longitudinal Study of Adolescent Health (ADD Health Study). The ADD Health study is a school-based investigation of the health-related behaviors of adolescents in junior high and high school. It was designed to explore the causes of these behaviors, with an emphasis on the influence of social context (i.e., the role families, friends, schools and communities play in the lives of adolescents that may encourage healthy or unhealthy behaviors). Data were collected in surveys of students, parents, and school administrators.

The general aim of this study is to examine the effects of and remedies for unit and item nonresponse in surveys that are longitudinal in nature. In doing so, we are investigating the effects of nonresponse on data gathering in the first two waves of the ADD Health Study. Data collection in the first wave includes a self-administered questionnaire completed by selected students in schools (IS1) and a subsequent in-person interview completed with the students' families in their homes (IH1). Families interviewed in Wave 1 were subsequently interviewed face to face in their home in a second round of data collection (IH2).

Our objectives apply to longitudinal methods in general and include: (i) developing expressions for total bias due to unit nonresponse—and its components—for round-specific estimates and in round-to-round differences for parameter estimates; (ii) estimating total and component biases—and identifying its determinants—for unit nonresponse in estimating key outcome measures; (iii) developing and evaluating alternative weighting adjustment strategies for round-specific unit nonresponse; and (iv) investigating the use of multiple imputation in dealing with item nonresponse.

To accomplish the first two objectives, we use an additive bias decomposition for the well-known Hansen-Hurwitz two-stratum model for nonresponse. Following Groves (1989) and Lessler and Kalsbeek (1992), the bias due to nonresponse (e.g., for an estimated mean) can be modeled as,

$$B = \lambda_{nr}\left(\bar{Y}_r - \bar{Y}_{nr}\right) = \sum_{c=1}^{4} B_c = \sum_{c=1}^{4} \lambda_{nr}^{(c)}\left(\bar{Y}_r - \bar{Y}_{nr}^{(c)}\right)$$

where λ_{nr} is the proportion of the population in the nonrespondent stratum, and $\bar{Y}_r$ and $\bar{Y}_{nr}$ are the means of those in the respondent and nonrespondent strata, respectively. Note also that $B_c = \lambda_{nr}^{(c)}\left(\bar{Y}_r - \bar{Y}_{nr}^{(c)}\right)$ is the component of bias due to the c-th type of nonresponse (e.g., noncontacts, refusals, etc.), where $\lambda_{nr}^{(c)}$ is the proportion of the population in the c-th nonrespondent substratum, and $\bar{Y}_{nr}^{(c)}$ is the mean for the same substratum.

Since substantive IS1 data are available for IH1 respondents and nonrespondents, and since IS1 and IH1 data are available for IH2 respondents and nonrespondents, this basic model can be used to estimate bias due to conditional nonresponse during the In-Home phases of data collection in the Add Health Study. We are also developing similar bias expressions in estimating cross-sectional change. Early findings applied to several measures of adolescent health risk, and included in the longer paper but not in this summary, generally find B<0 and components of bias to often be countervailing (e.g., $B_c > 0$ for refusals and $B_c < 0$ for other components). The implications of this type of "tug-of-war" among types of nonresponse suggest that one may need to be cautious in minimizing the extent of nonresponse due to certain types of nonresponse (e.g., refusals).

For the third objective, the team will develop a second set of weights for the Add Health data. Weights prepared by the data collection contractor (National Opinion Research Center) using the standard weighting class approach to adjust for nonresponse, will be compared to a nonresponse adjusted set of weights where the adjustment for nonresponse is a bounded estimate of the respondent's response propensity from a fitted logistic model. This methodology will be applied to cumulative Add Health nonresponse through the IH2 interview. Since data from the IS1 questionnaire will be available for most of the IH2 nonrespondents, we will be able to gauge the portion of IH2 cumulative nonresponse that is offset by the two adjustment strategies, thus providing a direct validation measure to assess the relative utility of the two approaches. This approach will build on prior work by Iannichione, Milne, and Folsom (1991) as well as Folsom and Witt (1994), the latter on weight adjustments for the Survey of Income and Program Participation.

The fourth objective will first involve an exploration of item nonresponse in the Add Health Study. Multiple imputation methodology will then be explored as a solution to remedy this level of nonresponse and measure its effects in the analysis of Add Health data. Work in this step of the study will first involve an exploration of item nonresponse in the Add Health Study. We will then profile the levels of item nonresponse on the study data files to identify the extent of the item nonresponse problem for Add Health and to thereby establish where imputation may have its greatest potential use to this study. Next, we will develop an appropriate plan for creating a data framework for the use of multiple imputation. The repeated imputation inferences are derived using a Bayesian paradigm, which requires the correctness of model specifications. Since this cannot be ascertained in practice, we will follow Rubin (1987) to evaluate this application of multiple imputation by adapting his randomization-based frequentist paradigm. This will be done to assess the sensitivity and robustness of multiple imputation to model deviations and finite number of imputations.

Conclusions

Each of the four research cores described above focuses on different substantive and methodological issues: (i) Study I explores the sampling and measurement issues involved in obtaining accurate

perinatal outcomes data among the Latino population of North Carolina; (ii) Study II advances file linkage technology to improve current methods for linking important medical databases (e.g., Medicaid files with hospital discharge data); (iii) Study III develops SAE methodology to identify "hot" or high risk areas at the county level; and (iv) Study IV explores way to deal with nonresponse in longitudinal research, particularly in the ADD Health Study.

Each of these studies is devoted to the integration of methodological advancement in the context of substantive research issues as it applies to high-risk populations in the area of public health. The sole purpose of our Center is to provide a vehicle or mechanism that allows this type of collaborative work to happen. Our hope is that such cross-fertilization will broaden the tools available to both the practice-oriented and research-oriented professionals in the field of public health.

REFERENCES:

Aneshensel, C. S., Bercerra, R. M., Fielder, E. P. & Schuler, R. H. (1989). Participation of mexican american female adolescents in a longitudinal panel survey. Public Opinion Quarterly, 53, 548-562.

Buekens, P., Notzon, F. Kotelchuck, M. & Wilcox, A. (2000). Why do mexican-americans have few low birth weight infants? American Journal of Epidemiology, 152, 347-351.

Copas, J. B., & Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records (with discussion). Journal of the Royal Statistical Society, 153(3), 287-320.

Elias-Olivares, L. & Farr, M. (1991). Sociolinguist analysis of mexican-american patterns of non-response to census questions (Ethnographic Exploratory Research Report # 16). Washington, DC: U.S. Bureau of the Census.

Fellegi, I. P., & Sutter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.

Folsom, R. E., Shah, B., & Vaish, A. (1999). Substance abuse in states: A methodological report on model based estimates from the 1994-1996 national household survey on drug abuse. Proceedings of the Survey Research Methods Section of the American Statistical Association, 371-375.

Folsom, R. E. & Witt, M. B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. Proceedings of the Survey Research Methods Section of the American Statistical Association , 428-433.

Forsyth, B. H. & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. A. Sudman (Eds.) Measurements Errors in Surveys (393-418). NY: Wiley & Sons.

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

Groves, R. M. (1989). Survey Errors and Survey Costs. New York: Wiley & Sons.

Guendelman, S. (1998). Health and disease among hispanics. In S. Loue (Ed.) Handbook of Immigrant Health (pp. 277-301). New York: Plenum Press.

Hines, A. M. (1993). Linking qualitative and quantitative methods in cross-cultural survey research: Techniques from cognitive science. American Journal of Community Psychology, 21(6), 729-746.

Iannacchione, V. G., Milne, J. G., & Folsom, R. E. (1991). Response probability weight adjustment using logistic regression. Proceedings of the American Statistical Association, 637-642.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association, 84, 414-420.

Kalsbeek, W. D. (1988). Design strategies for nonsedentary populations. Proceedings of the American Statistical Association, 28-37.

Kalsbeek, W. D., & Cohen, S. B. (1978). Disproportionate sampling in the national medical expenditure survey. Proceedings of the American Statistical Association, 276-281.

Lepkowski, J. M. (1991). Sampling the difficult-to-sample. Journal of Nutrition, 121(3), 416-423.

Lessler, J.T., & Kalsbeek, W. D. (1992). Nonsampling Errors in Surveys. New York: Wiley & Sons.

Malec, D., Sedransk, J., Moriarity, C. L., & LeClere, F. B. (1997). Small area inference for binary variable in the national health interview survey. Journal of the American Statistical Association, 92, 815-826.

McKay, R.B., & Aguirre, A. (1994). The Spanish Translation of the Redesigned Current Population Survey – Lessons Learned. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Danvers, Massachusetts.

Nandram, B., Sedransk, J., & Pickle, L. W. (in press). Bayesian analysis of mortality rates for u.s. health service areas. Sankhya, (Ser. B).

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley & Sons.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, CA: Jossey Bass.

Willis, G. B., Royston, P. & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. Applied Cognitive Psychology, 5, 251-267.